Schriften des Forschungszentrums Jülich
Reihe Schlüsseltechnologien / Key Technologies Band / Volume 113

Forschungszentrum Jülich GmbH
Institute for Advanced Simulation
Institute of Complex Systems
Jülich Centre for Neutron Science
Peter Grünberg Institute

Lecture Notes of the
47th IFF Spring School 2016

Rainer Waser, Matthias Wuttig (Eds.)

Memristive Phenomena –

From Fundamental Physics to Neuromorphic Computing

This Spring School was organized
by the Peter Grünberg Institute,
Forschungszentrum Jülich and
Physics Institutes, RWTH Aachen University,
Jülich Aachen Research Alliance,
Section Fundamentals of Future
Information Technology (JARA-FIT)
on 22 February – 4 March 2016.

In collaboration with
universities, research institutes and industry.

Schriften des Forschungszentrums Jülich
Reihe Schlüsseltechnologien / Key Technologies Band / Volume 113

ISSN 1866-1807 ISBN 978-3-95806-091-3

Bibliographic information published by the Deutsche Nationalbibliothek.
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Publisher: Forschungszentrum Jülich GmbH
IAS, ICS, JCNS, PGI
52425 Jülich
Phone +49 (0)2461 61-6048 · Fax +49 (0)2461 61-2410

Cover Design: Grafische Medien, Forschungszentrum Jülich GmbH

Printer: Schloemer + Partner GmbH, Düren

Distributor: Forschungszentrum Jülich GmbH
Zentralbibliothek, Verlag
52425 Jülich
Phone +49 (0)2461 61-5368 · Fax +49 (0)2461 61-6103
e-mail: zb-publikation@fz-juelich.de
Internet: http://www.fz-juelich.de

Schriften des Forschungszentrums Jülich
Reihe Schlüsseltechnologien / Key Technologies Band / Volume 113

ISSN 1866-1807
ISBN 978-3-95806-091-3

Contents

Preface

Introduction and Survey
Rainer Waser

A Fundamentals

A1 Structure of Matter – From Perfect Crystals to Amorphous Materials
David P. DiVincenzo

A2 Electronic Structure of Matter
Stefan Blügel and Gustav Bihlmayer

A3 Lattice disorder in ionic crystals
Felix Gunkel

A4 Ion Transport in Metal Oxides
Roger A. De Souza

A5 Phase Transitions
Christian Pithan

A6 Physics and Chemistry of Redox Processes
Rotraut Merkle

A7 Electron Transport: Disorder and Correlations
Matthias Wuttig

A8 Magnetism and Spin-Polarized Transport
Daniel E. Bürgler

A9 Electron Tunneling
Daniel Wortmann, Phivos Mavropoulos

B Technology

B1 Chemical Vapour Deposition Techniques
Susanne Hoffmann-Eifert

B2 Physical Deposition Techniques
Regina Dittmann

B3 Nanotechnological Integration
Jürgen Moers

B4 Self-Organization Techniques
Bert Voigtländer

C Analysis and Characterization

C1 Electrical Characterization of Memristive Cells
Ulrich Böttger and Viktor Havel

C2 X-Ray Diffraction and Scattering
Uwe Klemradt

C3 From Atomic Structure to Properties of Oxides – Applications of Aberration-corrected TEM
Chun-Lin Jia

C4 HRTEM Based Spectroscopy Techniques
Christopher Brian Boothroyd

C5 Photoelectron Spectroscopy
Lukasz Plucinski

C6 Electron Emission and Photoemission Microscopy
Claus M. Schneider

C7 Scanning Probe Microscopy
Philip Ebert, Marco Moors

D Memristive Phenomena for Non-Volatile Electronic Functions

D1 Magnetic Random Access Memory
Ioan Lucian Prejbeanu

D2 Electrochemical Metallization Memories
Ilia Valov

D3 Valence change in Nanoionic Oxide Cells
Regina Dittmann

D4 Switching Kinetics of Redox-based Resistive Memories
Stephan Menzel

D5 Electronic Avalanche in Narrow Gap Mott Insulators and Non-Volatile Memories
Etienne Janod, Benoit Corraze, Julien Tranchant, Marie-Paule Besland and Laurant Cario

D6 Phase Change Materials
Matthias Wuttig

D7 Threshold and Memory Switching Kinetics of Phase Change Materials
Martin Salinga

D8 Interfacial Phase Change Materials
Riccardo Mazzarello

D9 Memristive Tunneling Devices: From Device Principles to Neuromorphic Applications
Martin Ziegler, Adrian Petraru, Rohit Soni and Hermann Kohlstedt

E Applications and Future Directions

E1 Reliability of Memristive Elements
Arne Heittmann, Tobias G. Noll, Dirk J. Wouters, Yang-Yin Chen Andrea Fantini, Nagarajan Raghavan

E2 Ultimate Physical Limit of Scaling
Victor V. Zhirnov

E3 Select Devices For Memristive Crossbar Arrays
Dirk J. Wouters

E4 From Memristive Gate-Array Logic to Neuromorphic Computing
Eike Linn and Arne Heittmann

Appendix

Preface

Memristive phenomena combine the functionalities of electronic resistance and data memory in solid-state elements, which are able to change their resistance as a result of an electrical stimulation in a non-volatile fashion. In nanoelectronics, this functionality can be used for information storage and unconventional logic, as well as neuromorphic computing concepts that are aimed at mimicking the operation of the human brain.

A multitude of fascinating memristive phenomena has emerged over the past two decades. These phenomena typically occur in oxides and higher chalcogenides and are one of the hottest topics in current solid-state research, comprising unusual phase transitions, spintronic and multiferroic tunneling effects, as well as nanoscale redox processes by local ion motion. They involve electron correlation, quantum point contact effects and exotic conformation changes at the atomic level.

The Spring School provides a comprehensive introduction to and an overview of current research topics covering the physics of memristive phenomena, with an emphasis on an understanding of the underlying basic principles. The inspiration to organize this school arose from our Cooperative Research Center **Resistively Switching Chalcogenides for Future Electronics (SFB 917)** which has been funded by the Deutsche Forschungsgemeinschaft since July 2011. The overarching aim of the SFB 917 is to advance the microscopic understanding of memristive phenomena utilizing changes in the atomic configuration, in particular in the phase and the valence of oxides and higher chalcogenides. To explore the full potential and pave the way for an ultimately energy-efficient electronics technology it is mandatory to realise ultrahigh scalability, fast switching kinetics and long retention times. The promise to realise fast, non-volatile devices which may enable novel, brain-like functionalities by neuromorphic computing, defines the technological potential of the SFB.

The school comprises approximately 50 hours of lectures, including discussions, as well as the opportunity to visit the participating Institutes in Forschungszentrum Jülich. All lectures will be given in English. Registered participants will receive a book of lecture notes that contains all of the material presented during the school. The lectures are grouped together in five sections, which are outlined below.

Fundamentals

Material properties provide the basis for understanding the physics of the processes that occur during memristive phenomena. These lectures focus on atomic and electronic structure, with an emphasis on metal oxides and higher chalcogenides that are used in memristive cells, lattice disorder in these materials, phase transition processes and ionic transport mechanisms. This section concludes with the physics of redox processes, electron transfer at interfaces, electronic transport properties including correlation effects, insulator-metal transitions, and electron tunneling.

Technology

In order to achieve functional and energy-efficient electronic circuits, memristive cells must be integrated with CMOS digital circuits using process technology and designed to match specific

applications, such as information storage in memory or information processing in logic circuits. Cell design and state-of-the-art integration technology are introduced in this section of the Spring School. The lectures focus on chemical and physical vapor deposition techniques (molecular beam epitaxy, pulsed laser deposition and sputtering), in combination with optical, electron-beam or nano-imprint lithography, ion etching and atomic precision polishing. As an alternative to these top-down techniques, promising bottom-up approaches that involve molecular self-assembly processes are described briefly.

Analysis and characterization

Functional characterization of memristive cells can be performed by using electric sweep and pulse tests over a wide dynamical range, from sub-nanoseconds to kiloseconds. In order to understand the microscopic mechanism of memristive phenomena, a plethora of advanced microscopic, scattering, and spectroscopic techniques are required. Due to the fact that memristive cells are typically nanoscale objects and atomic and electronic configuration changes that lead to resistance changes are tiny, the elucidation of the operating principles of memristive cells is highly challenging and sometimes beyond the possibilities of the techniques that are currently available. These lectures cover X-ray diffraction methods, aberration-corrected high-resolution transmission electron microscopy and spectroscopy, off-axis electron holography and tomography, photoemission electron spectroscopy and microscopy, and cutting-edge scanning probe techniques.

Memristive phenomena for non-volatile electronic functions

The fascinating internal physical mechanisms of memristive phenomena in oxides and higher chalcogenides can be grouped into nanoscale phase transitions, nanoionic redox processes and the modulation of the tunnel transmission through barriers because of changes in the barrier or the terminals. Magnetic terminals lead to the chance to exploit magnetoresistive and spintronic effects, such as spin-transfer torque. These lectures will cover nanoionic redox processes, *i.e.*, processes in which local ion motion in metal oxides is highly non-linear as a result of thermal and/or field enhancement, leading to a valence change of the metal ions and a corresponding modulation of the electron transport. Phase change memories rely on volatile electronic threshold switching followed by a thermally-induced phase transition between an amorphous and a crystalline state, in which disorder controls electron transport. In the case of functional tunneling oxides, electron transmission can be modulated by electrostatic, ferroelectric, multiferroic or nanoionic effects in the tunneling barrier.

Applications and Future Directions

The last section of the Spring School comprises lectures that cover present and future application areas, such as memories for information storage, unconventional logic and neuromorphic computing concepts that are aimed at emulating the function of the human brain. Aspects of memristive circuits such as the required selector devices in array architectures, as well as reliability issues and ultimate physical limits of further miniaturization, will be explained.

This school could not take place without the help and dedication of many colleagues. We are grateful to all contributors from the Peter Grünberg Institute (PGI) of the Forschungszentrum Jülich and colleagues from the RWTH Aachen University as part of the Jülich-Aachen Research Alliance, section Fundamentals of Future Information Technology (JARA-FIT).Explicitely, we acknowledge the time and effort the following colleagues spent to prepare the manuscripts and the lectures:

Dr. Gustav Bihlmayer (PGI-1/IAS-1, FZJ)
Prof. Stefan Blügel (PGI-1/IAS-1, FZJ)
Dr. Chris Boothroyd (PGI-5/ER-C, FZJ)
Dr. Ulrich Böttger (IWE2, RWTH)
Dr. Daniel Bürgler (PGI-6, FZJ)
Dr. Regina Dittmann (PGI-7, FZJ)
Prof. Dunin-Borkowski (PGI-5/ER-C, FZJ)
Dr. Philip Ebert (PGI-5, FZJ)
Dr. Felix Gunkel (IWE2, RWTH)
Victor Havel (IWE2, RWTH)
Dr. Arne Heittmann (EECS, RWTH)
Dr. Susanne Hoffmann-Eifert (PGI-7, FZJ)
Dr. Chun-Lin Jia (PGI-5/ER-C, FZJ)
Prof. Uwe Klemradt (Physics 2C, RWTH)
Dr. Eike Linn (IWE2, RWTH)
Dr. Phivos Mavropoulos (PGI-1/IAS-1, FZJ)
Dr. Riccardo Mazzarello (ITSSP, RWTH)
Dr. Stephan Menzel (PGI-7, FZJ)
Dr. Jürgen Moers (PGI-8, FZJ)
Dr. Marco Moors (PGI-7, FZJ)
Prof. Tobias Noll (EECS, RWTH)
Dr. Christian Pithan (PGI-7, FZJ)
Dr. Lukasz Plucinski (PGI-6, FZJ)
Dr. Martin Salinga (Physics IA, RWTH)
Prof. Claus M. Schneider (PGI-6, FZJ)
Dr. Roger De Souza (IPC, RWTH)
Dr. Ilia Valov (PGI-7, FZJ)
Prof. David DiVincenzo (PGI-2, FZJ)
Prof. Bert Voigtländer (PGI-3, FZJ)
Dr. Daniel Wortmann (PGI-1/IAS-1, FZJ)
Dr. Dirk Wouters (IWE2, RWTH)

We highly appreciate that several distinguished colleagues from external universities and research laboratories have agreed to contribute to the program of the school:

Dr. Yang-Yin Chen, Dr. Andrea Fantini, imec, Belgium

Dr. Etienne Janod, Dr. Benoit Corraze, Dr. Julien Tranchant, Dr. Marie-Paule Besland and Dr. Laurent Cario, Institute des Matériaux Jean Rouxel (IMN), CNRS Nantes, France

Dr. Rotraut Merkle, Max-Planck-Institut für Festkörperforschung, Stuttgart

Dr. Ioan Lucien Prejbeanu, SPINTEC, UMR 8191 CEA-CNRS-UGA-GINP, Grenoble, France

Dr.Nagarajan Raghavan, imec, Beligum & Singapore Univ.

Prof. Victor Zhirnov, Semiconductor Research Corporation (SRC) and NCSU, USA

Dr. Martin Ziegler, Dr. Adrian Petraru, Dr. Rohit Soni, Prof. Hermann Kohlstedt, Nanoelektronik, Christian-Albrecht-Universität zu Kiel

We would like to express our thanks to all these colleagues for the effort and enthusiasm, which they have put into the preparation and presentation of their lectures and manuscripts. We are very grateful to the board of directors of the Forschungszentrum Jülich for the continuous organizational and financial support, which we have received for the realization of the IFF Spring School and for the production of this book of lecture notes. Furthermore, our special thanks go to Michael Beissel for the general management, to Thomas Pössinger for the preparation of the layout, Dagmar Leisten for graphics support, as well as Maria Garcia and Luise Snyders for countless supporting tasks.

Finally, we are highly indebted to the Deutsche Forschungsgemeinschaft for the support of our Cooperative Research Center (SFB 917).

Jülich and Aachen, January 2016

Rainer Waser and Matthias Wuttig
School Chairmen

Introduction and Survey

Rainer Waser

Peter Grünberg Institut, PGI-7, Forschungszentrum Jülich GmbH

Institut für Werkstoffe der Elektrotechnik II, IWE2
RWTH-Aachen University

JARA-FIT

Contents

1 Binary switch – the most basic element representing information 2

2 Conventional nanoelectronic devices and their physical limits 4

3 Concept of two-terminal memristive elements 8

3.1 Advantage of resistive switching over conventional non-volatile concepts 8

3.2 Operation modes of memristive devices 9

3.3 Classification based on memristive mechanisms 11

3.4 Memristive Systems and Memristors 18

4 Prospects and Challenges 19

This chapter provides a brief sketch of the framework in information theory and physics for the Spring School. We will start with physical state variables which are used to represent information, basic definitions of memristive switching elements, their main switching modes, and their most important performance parameters. Furthermore, it sketches the scope of the book which spans from nanoscale physics and chemistry of the switching phenomenon to devices, technology, and application areas. A classification of memristive elements and a brief history of the phenomena is given.

This chapter comprises parts from the following sources: General Introduction and Introduction to Part III and Part V of "Nanoelectronics and Information Technology" [1], Chapter 4 of "Nanotechnology – Information Technology I" [2], Chapters 3-9 of "Emerging Nanoelectronic Devices" [3], and Chapter 1 of "Resistive Switching" [4].

1 Binary switch – the most basic element representing information

Information is coded in languages. The smallest, i.e., irreducible element of a language are called **characters**. The basic set of characters of a language constitutes the **alphabet** or code of this language. For instance, the Morse alphabet consists of three discrete basic characters, that is dot, dash, and space. The English language exhibits a character set of 26 letters, in upper- and lower case, ten numerical digits, and various punctuation marks. The genetic DNA code uses four discrete characters represented by four different chemical base groups. The binary code, or **Boolean code**, consists of only two characters, "0" and "1". Since any language requires at least two discrete basic characters in order to represent information, the Boolean code is the most elementary language. This makes it most suitable for carriers with bi-stable states, frequently used in digital electronics which is the basis of the **Information and Communication Technologies (ICT)** today. The information I encoded in one character of the Boolean code is called 1 **bit**.

In the context of information processing, logical operations are applied at the basic characters of a language. This is why the basic characters are also called **logic states**. Digital electronics today relies on binary Boolean logic, although multinary logic concepts are conceivable as well and will be discussed in the course of this Spring School.

In order to be transmitted, processed or stored, information always requires a physical carrier. The information is mapped onto physical properties, the so-called **state variables**, of the carrier by structuring or patterning the carrier in space or time. The smallest binary information systems consists of a **binary switch**. An abstract binary switch is sketched in Fig. 1 illustrating the basic operations WRITE, READ, and TALK. The WRITE operation sets a dedicated state in the binary switch, the READ operation detects the state of the switch, and the TALK operation transfers the state from/to the switch (without detecting the state). These basic operations can be used to cover the three areas of information technology: **information storage** by writing and reading the memory, **information processing** by logic operations, and **information transfer** by communicating the information.

As expressed by the terms **electronics, microelectronics,** and **nanoelectronics**, the input and output signals to binary switches today is based on electrons as the physical carriers, using the

electron charge (as oppose to its spin or mass) as the state variable, although other possibilities exist too (e.g. concentration differences of ions in biological neurons, the phase of Cooper pairs in superconductor logic circuits, etc. [1]). *Within* the binary switch, various state variables may be used. A main goal of research in emerging concepts of nanoelectronics is the investigation of these state variables and the corresponding physical carrier.

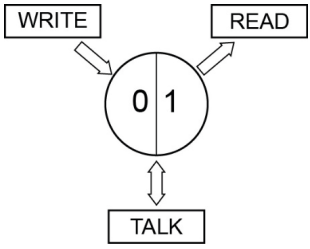


Figure 1: *The constituents of an abstract binary switch (From Chap. 4 in Ref. [2]).*

Fig. 2 shows the major options for state variables and corresponding physical carriers reported in literature. They are not completely orthogonal, i. e., different properties of a carrier can be used as state variable. For example, the charge or the magnetic dipole orientation (spin orientation) of an electron can be used as the state variable for computation. Furthermore, the spin of electrons in nanodots may be utilized by the phase of their wave functions in order to create magnetic qubits. Often, a combination of effects has to be considered. For example, an electronic charge transferred to and localized on an atom in molecules may lead to a conformation change (i. e., a change in the geometrical arrangement of the atoms in the molecule) which in turn may result in a stabilization of the state (Chap. 21 in [1]). Here, the electron charge *or* the arrangement of atoms may be considered as the state variable. The same holds for, e. g., for some redox-based resistive memory cells were the depletion of anions, i. e., a rearrangement of atoms, leads to a change in the oxidation states of cations i. e. the localized electron charges, as we will discuss in details.

Another classification of (binary or multinary) switches refers to the number of terminals:

- **two-terminal devices** in which the same terminals are used for the input signal to modify (write) the output state and for the reading of the output signal (as well as for the energy supply); examples are: diodes, capacitors, and, in particular, **memristive elements** as the topic of this Spring School.
- **three-terminal devices** (in general: multi-terminal devices) in which the input (control) signal uses a separate terminal than the output signal; example: transistors. In transistors, the input signal is applied to the gate (or base) while the output signal is the electronic current from source to drain (or emitter to collector).

Research in nanoelectronics mainly focusses three overarching aims: (1) increase of the computational throughput in information processing by alternative concepts, (2) reduction of the energy consumption by orders of magnitude and exploiting the physical limits in energy-efficiency, and (3) improvement in the device density by further miniaturization because of costs and because of new areas of applications (such as the Internet of Things, IoT). Memristive phenomenon may hold the key to all three of these aims. Memristive elements are two-terminal devices which inherently decreases the wiring complexity on chips significantly because a separate wiring for the control signals is not required. As we shall see in the following chapters, memristive phenomena may also lead to energy efficiencies much beyond the limits of conventional nanoelectronics, and they hold the promise of a paradigm shift into completely new computational concepts.

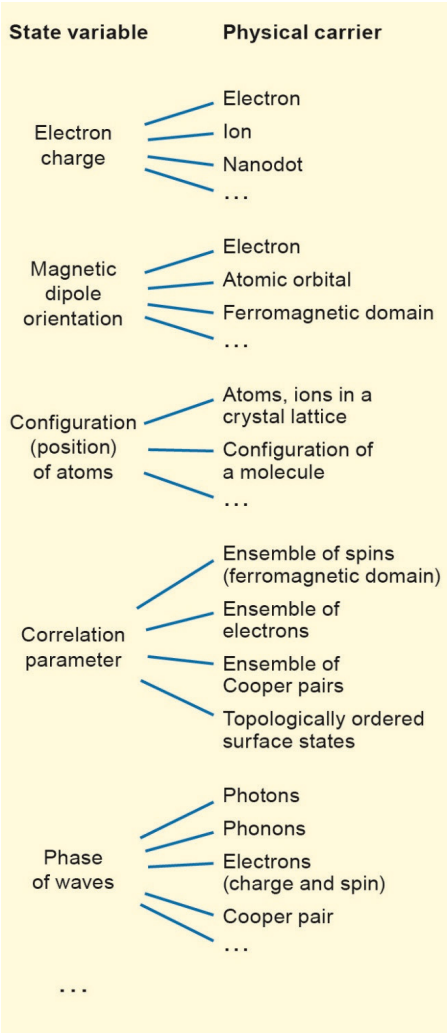


Figure 2: *Options for state variables and corresponding physical carriers. (From Introduction to Part III in Ref. [1], with modifications).*

2 Conventional nanoelectronic devices and their physical limits

In order to demonstrate the prospects of devices based on memristive phenomena, we will briefly discuss few prominent examples of nanoelectronic devices and their physical limits upon further miniaturization. We will choose the most common memory devices used today, DRAM and Flash, because of their ubiquitous use and because of their memory functionality which we are going to compare with memristive devices throughout this Spring School.

As a measure of the miniaturization we shall use the **minimum feature size *F*** which denotes the smallest feature (in nm) which can be fabricated by a given technology of a semiconductor factory. *F* is mainly determined by the lithography used for fabricating the integrated circuits. The reduction of *F* over the years is empirically reflected by the famous Moore´s law. *F* has been 1 µm in 1988, 130 nm in 2002, and it is currently (2016) 14 nm for the most advanced processor and memory chips in mass fabrication by companies such as Intel, Samsung, and TSMC.

Dynamic random access memory (DRAM) is the most prominent operation memory in computer systems today – from smart phones and tablets to servers and super computers. The storage element is a dielectric capacitor as a two-terminal device operating on electron charge as the state variable. The direction of the charging (one capacitor plate positively charged and the other one negatively charge, or vice versa) represents the binary state, "0" or "1". In order to keep the charge on the capacitor plates, the capacitor is connected to an access transistor T which isolates the storage capacitor unless there is a READ or WRITE operation. These combinations of a storage capacitors and access transistors (short: 1T-1C) are the memory cells of the DRAM and they are organized in an array (or: matrix) configuration as in every random access memory. An array with access (or: select) transistors at each node is called an **active array**, while an array with only the two-terminal storage elements is called a **passive array** (Fig. 3). In the context of arrays of memristive devices, this issue will be discussed in E3.

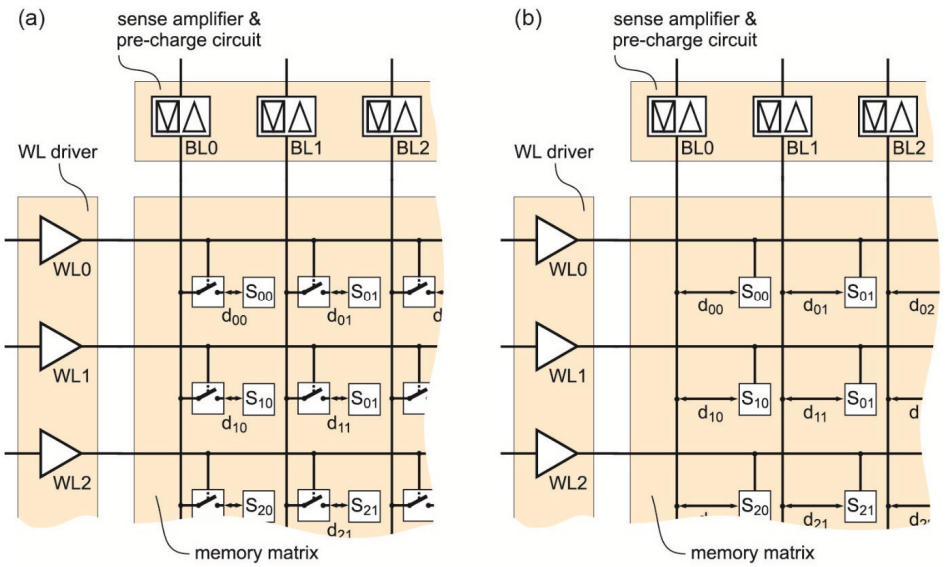


Figure 3: *Configuration of array-based memories including the word line (WL) drivers located at the rows of the array and sense amplifier / pre-charge circuit units driving and sensing the bit lines (BL) located at the columns of the array. The sketch is generalized and simplified. The actual configuration will depend on the type of memory and design.*
(a) Active array.
(b) Passive array.
S_{ik} storage elements, d_{ik} data signals
In the case of DRAM, the storage elements are dielectric capacitors.
(From Introduction to Part V in Ref. [1])

In the active array, the word lines (WL) are connected to the gates of the transistors, while the bit lines (BL) are connected to the storage element. In the case of the DRAM, the BL accesses one of the plates of the storage capacitor (with a capacitance C_S) via the channel of the transistor (Fig. 4). The other plate of the storage capacitor is connected to a plate line (PL) which is typically kept at a fixed voltage (e.g. 1/2 of the operating voltage). The DRAM cell has only a short time for which it can keep the information. The **retention time** is less than 1 s because of leakage currents through the capacitor dielectrics and leakage through the transistor despite being turn off. For this reason, the charge state of the storage capacitor is periodically refreshed, e.g. every 64 ms. Because of the low retention time and the fact that the information is lost when the supply voltage of the circuit is turned off, DRAM belong to the group of **volatile memories**.

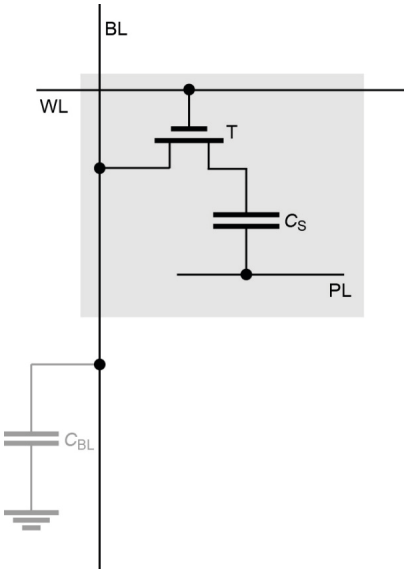


Figure 4: *Schematic representation of a DRAM cell as part of an array.*

For a WRITE operation, the pre-charge circuit sets the required voltage to the BL of the cell to be addressed. For a READ operation, the charge of the addressed storage capacitor is redistributed between C_S and the parasitic BL capacitance C_{BL} as in a capacitive voltage divider. Because an array consists of, for instance, 512 WL and 512 BL, the C_{BL} is quite high, e.g. 100 fF for a DRAM in the 45 nm technology; see Chap. 27 in Ref. [1]. Typically C_S should not be less 20 fF in this case, in order to cover the sensitivity of the sense amplifier at the end of the BL. To realize such relatively large capacitance values on a small footprint of an integrated chips (e.g. 6 F^2), the storage capacitor is put upright and extended into the third dimension either as a deep trench into the substrate of the chip or as a high stack above it. However, this concept runs into ultimate physical scaling limits because of the dimensions of the capacitor stack given by the sum of the dielectric material and the electrode thickness. The leakage currents become unacceptable large for high-permittivity dielectrics below a dielectric thickness of approx. 8 nm, and the electrode thickness need to be at least 2 nm in order to obtain sufficiently high conductances. For these reasons, the ultimate scaling limit is estimated at F = 12 nm, while technological limits will set in even before. At this limit, the real area of the storage capacitor (which is folded into the trench or stack) is >300 F^2, i. e., approx. 40000 nm^2. Details can be found in Chap. 27 in Ref. [1].

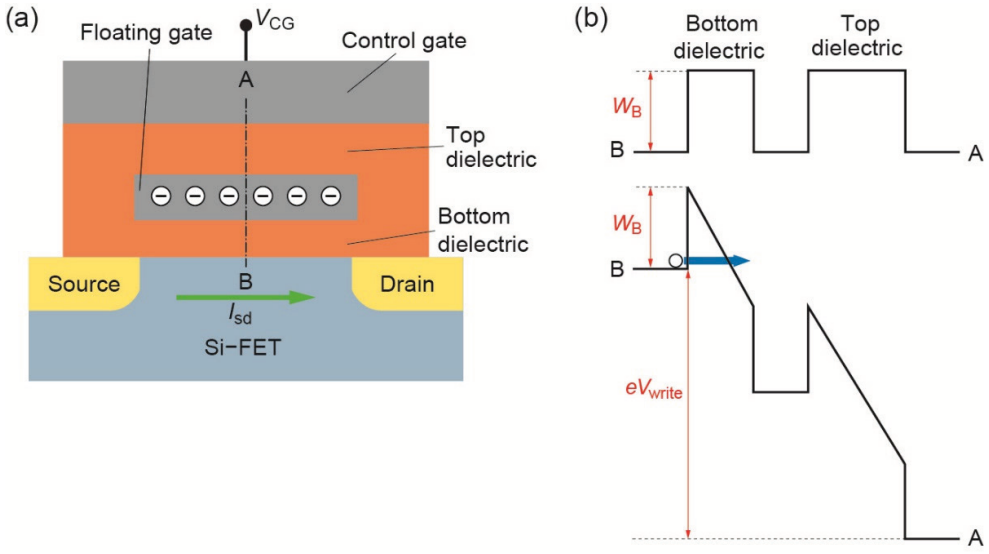


Figure 5: *(a) Cross section of a floating gate MOSFET. (b) Electrostatic energy barrier diagram from the channel to the control gate (along cut A to B in (a)).*

Flash memories are based on MOSFETs (Metal-Oxide-Semiconductor Field-Effect Transistors) with an additional **floating gate**, invented by Sze and Kahng in 1967 [5]. The floating gate is an isolated gate embedded in the gate dielectrics of the MOSFET (Fig. 5).

If excess electrons are trapped on the floating gate, the threshold voltage of the transistor is shifted. This shift can be detected so that the floating gate MOSFET represents a memory element. If the energy barrier W_B established by the dielectrics (e.g. SiO_2) is high enough (> 1.7 eV) and the barrier is thick enough (> 5 nm) to prevent direct tunneling, the trapped excess electrons may be kept on the floating gate for more than 10 years which is a typical characteristic of a **non-volatile** memory.

The WRITE operation of a floating gate MOSFET is facilitated by high voltages applied to the control gate. At sufficiently high voltages, the dielectric barrier between the channel and the floating gate becomes so strongly deformed that electron tunneling through the trapezoidal part of the barrier (so-called Fowler-Nordheim tunneling) is possible (Fig. 5b). For a reasonably fast write operation (ms to µs), write voltages of > 12 – 15 V are required. The relationship between the write speed and the write voltage represents the voltage-time dilemma, which is a fundamental property of non-volatile memories: if one wants to reduce the operating voltage, a dramatic (approx. exponential) degradation of the operation speed results. As an alternative write operation, injection of hot electron from the channel into the floating gate can be used. Relatively high voltages between source and drain and at the control gate are required and the process is very inefficient, as only one electron out of 10^5 to 10^6 electrons gets injected.

Highly dense arrays of floating gate MOSFETs are called Flash memories, in which the information can only be erased in blocks of e.g. 1 MB. Today, this is by far the most ubiquitous type of integrated non-volatile memories, and it is used in USB memory sticks, smart phones, cameras, and increasingly solid-state drives (SSD) replacing magnetic hard disc drives in laptop computers, serves, and super computers.

Despite the huge market penetration today, severe scaling limits have been reached already. The thickness of the stack from the channel to the control gate is inherently limited by direct tunneling, the voltages are extremely high compared to operating voltages in modern microprocessors, and the devices cannot be individually addressed in the high dense Flash architectures (i.e. there is no random access). Details about floating gate MOSFETs and Flash memories can be found in Chap. 26 in Ref. [1].

3 Concept of two-terminal memristive elements

3.1 Advantage of resistive switching over conventional non-volatile concepts

As we have discussed for the DRAM cell, capacitors require a relatively large area in order to build up a significant capacitance value (to compete with any BL capacitance). More importantly, the cross section of a resistor may be extremely small. In contrast, the cross section of a resistor may be tiny. Even a chain of single metal atoms may exhibit reasonable resistance value in the low kΩ range. Such a chain would have a cross section of < 1 nm^2 (compared to approx. 40000 nm^2 for a DRAM capacitor at the scaling limit). Taking out just two metals atoms (or changing them into atoms of insulators) from this chain increases its resistance by more than a factor of hundred.

In addition, as we have learnt from floating gate cells, confining an electron in order to use it for a non-volatile memory requires relatively high (W_B) and thick (a) energy barriers because of the low mass m of electron and the resulting high tunneling probabilities. If a non-volatile state is established by a **configuration of atoms**, the requirements on the surrounding energy barriers disappear because of the several 1000 times higher mass of atoms and the fact that the mass enters into the exponent of the tunneling probability:

$$\Pi \propto \exp\left(-a \cdot \sqrt{W_B} \cdot \frac{2\sqrt{2m}}{h}\right)$$

In fact, an atom even does not tunnel from its site in a crystal lattice to a neighbouring (empty) site.

For these two main reasons, any concept which relies on the change of the resistive on the nanoscale by changing the configuration of atoms (and some other means, as we shall see) may have a huge advantage over conventional non-volatile device concepts with respect to further miniaturization, i.e. the ultimate limits of scaling, and data retention. Concepts of resistance change to realize a binary (or multinary) switch are called, as mentioned, **resistive switching**. They denote reversible phenomena of 2-terminal elements which change their resistance upon electrical stimuli in a non-volatile fashion [6]. The *reversibility* is obtained by repeated applications of suitable stimuli which control the resistance value between two or more levels. *Non-volatility* means that the resistance change remains for a (long) retention time after the stimulus has been released. Phenomenologically, the stimulus affects an internal *state variable* of the element which controls the resistance. For this reason, the resistance values are memorized by the element which are, therefore, called **memristive** elements or devices (see Sec. 3.4 and Chapter E4). The required switching speed and the retention times depend on the area of application and will be discussed later. A memristive element always shows some kind of a MIM structure, composed of a more insulating (usually still well conducting) material I sandwiched between two (possibly different) electron conductors M as electrodes.

For any discussion of the stability of resistive states and data retention, one should keep in mind that only *one* of the states, ON state or OFF state or any intermediate state, can be **thermodynamically stable** (i.e. show the lowest free energy of the system), if the internal state variable is a scalar (such as a charge, a phase or a concentration). Inherently, the other state(s) must be **metastable.** If the internal state variable is a vector (such as a magnetization), the different resistive states can be represented by different directions of this vector – at the same energy. In any case, there must be a sufficiently high energy barrier between the states in order to allow for data retention.

3.2 Operation modes of memristive devices

Depending on the specific type of memristive device, different operation modes have to be used. Figure 6 shows schematically characteristic current-voltage (*I*-*V*) diagrams recorded by periodic voltage sweeps (left) and pulse sequences with voltage pulse excitation and current responses (right). By far the most device applications will use the **pulse mode**. However, the *I*-*V* **sweep mode** is helpful for obtaining an overview of the characteristics.

We will use the following notation for the states and the processes in memristive elements. The resistance states of a memristive cell are called High Resistance State (**HRS**) or **OFF state** and Low Resistance State (**LRS**) or **ON state**. For multilevel operation, intermediate resistance states are utilized as well. We assign the logic ´0´ state to the HRS and the logic ´1´ state to the LRS. A **write operation** changing a memristive cell from the HRS to the LRS is called a **SET** operation, while an opposite write operation is called a **RESET** operation.

Many memristive systems reported in the literature are operated in the **bipolar resistive switching (BRS** or **BS)** mode (Fig. 6a). Starting in the HRS, a SET process can be triggered by a voltage $V_{SET} > V_{th1}$ and leads to the LRS. Often a current compliance (cc) is used for the SET operation in order to avoid damage to the cell and to optimize the operation. A read operation is performed at a much smaller voltage magnitude V_{rd} to detect the current while avoiding a detectable change of the state. A voltage signal V_{RES} of opposite polarity and an amplitude $V_{RES} < V_{th2}$ is used for the RESET process to switch the cell back into the HRS.

The **unipolar resistive switching (URS** or **US)** mode (Fig. 6b) is characterized by the fact that all write and read operations can be performed with only one voltage polarity. For example, starting in the HRS, the SET process takes place at a voltage $V_{SET} > V_{th1}$, with a LRS current limited by a current compliance (cc). It is important that the cc is released in the RESET process with $V_{RES} > V_{th2}$, so that the current can exceed the cc value which leads to change back into the HRS. The read operation is performed at a small voltage V_{rd} as in the bipolar mode. For phase change memories (PCM), the operation is slightly different. Here, a long pulse with a moderate amplitude is used to SET the cell and a short pulse with a high amplitude is used to RESET the cell.

The **complementary resistive switching (CRS** or **CS)** mode (Fig. 1c) can be obtained by connecting two BRS-type memristive cells in an antiserial manner as suggested by Linn et al. [7]. Typically, the state of a CRS cell cannot be read at small voltages because the cell then always appears to be in a HRS. The state of the cell is only recognized at voltages $V > V_{th1}$. A read voltage $V_{rd} > V_{th1}$ will lead to a higher current (upper *I*-*V* trace in Fig. 1c, left) in the case of a logic ´1´ state, and to a lower current (lower *I*-*V* trace in Fig. 1c, left) in the case of a logic ´0´ state. The write ´0´ is achieved by a positive voltage $V_{wr} > V_{th2}$, and a write ´1´ is obtained by a negative voltage $V_{wr} < V_{th4}$. Because of the relative high read voltage amplitude $V_{rd} > V_{th1}$ and the corresponding currents, the internal state is affected by the read operation, i. e., the read voltage may destroy the logic state (so-called Destructive Read-Out, DRO). As a consequence, the last logic state needs to be re-written into the cell after every read operation. This is the same situation as in the case of the standard DRAM cells.

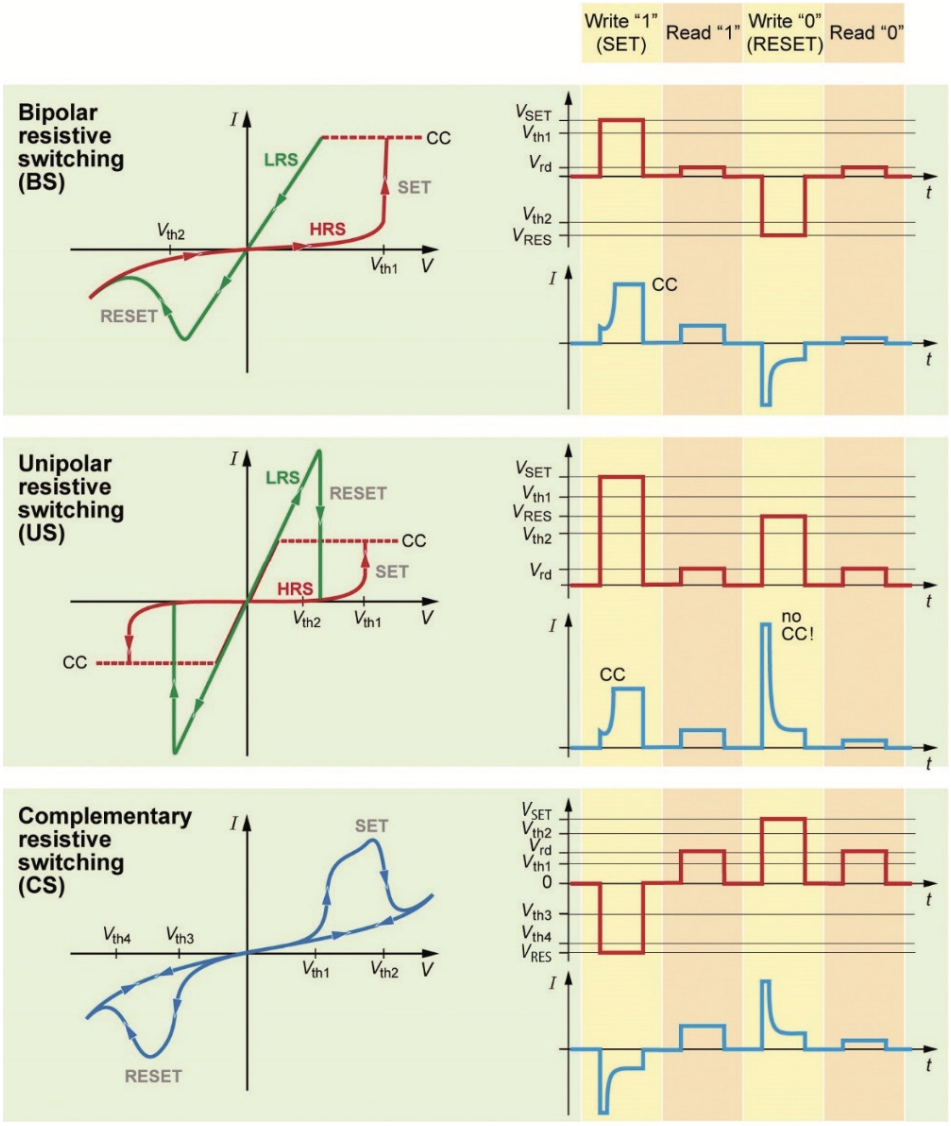


Figure 6: *The three most common operation modes of different types of memristive elements shown for the I-V sweep operation (left) and the pulse operation (right). Details are described in the text. Please note that the elements are non-volatile. At first glance, the CS (CRS) mode resembles the so-called threshold switching which shows a hysteresis above a certain voltage bias but which disappears at voltages below this bias. The difference is the fact that the information is lost in the case of a threshold bias while it is maintained in a CRS cell and can be read-out in the indicated manner (From [8], to be published).*

The major performance parameters of memristive device are:

- **Resistance values** R_{LRS} and R_{HRS} (or: R_{ON} and R_{OFF}), and the resistance ratio R_{HRS}/R_{LRS}.
- **SET and RESET voltages**, V_{SET} and V_{RES}, respectively.
- **Write currents**, in particular the current in the ON state, at a voltage amplitude just above V_{SET}.
- **Write speed** - the shortest electrical pulse able to change the resistive state.
- **Retention time** - the time for which a resistive state is maintained without a voltage applied to the cell.
- **Endurance** - the number of switching cycles before the resistance ratio fatigues to an inacceptable value.
- **Operation energy per bit** - the energy required to write a cell, i.e., to change its resistive state.
- **Scalability** - the geometrical size to which a cell can be miniaturized before it encounters inherent (physical) limits.
- **Stackability** - the option to stack several layers of cell on top of one another by fabrication technology.
- **Multilevel storage** – the option to store more than one bit of information in one cell.

3.3 Classification based on memristive mechanisms

The fundamental physical principles of memristive phenomena and, hence, the nature of the internal state variable can be manifold. In a coarse-grained classification, one can distinguish primarily between magnetic effects, electrostatic effects, and various classes of effects based on atomic configuration. Memristive phenomena have been investigated since the 1960s, and with a strong increase of the research activities in the last 20 years. The degree of understanding the mechanism, the potential with respect to miniaturization (scaling potential), efficiency in the switching energy, switching speed, and data retention varies considerably. In addition, clear scientific facts and pure speculations are often not well separated. Furthermore, different notations are used for the same effect. All these aspects contribute to the difficulty of an unequivocal classification of memristive phenomena. We propose such a classification in Figure 7, being aware that there are alternatives to this version. Within this Figure and throughout the Spring School book we place an emphasis on those phenomena for which the basic mechanism is reasonably well understood, which show a high potential for future nanoelectronics with respect to scaling, energy-efficiency, speed, and retention, and for which a reasonably large body of literature exists.

Magnetic effects

Concepts of spinelectronics (short: spintronics) including magnetic memristive effects emerge from the discovery of the giant magnetoresistance (GMR) effect by Peter Grünberg and Albert Fert in 1988. The GMR effect gives rise to a relatively large change in the resistance of a stack of a thin non-magnetic metal sandwiched between two magnetic metal layers, depending on the mutual relation of the direction of the magnetization of these layers. The GMR effect is exploited in spin-valves introduced by Dieny et al. in 1991 [9] which can be regarded as the prototype of magnetoresistance random access memories (MRAM). In 1995, the tunnel magnetoresistance

(TMR) in magnetic tunnel junctions (MTJ) was observed at room temperature [10, 11]. In a MTJ, a tunneling barrier, typically an insulating oxide, is sandwiched between two magnetic metals, one with a fixed magnetization direction and the other with the opportunity to change the magnetization direction by a moderate magnetic field. These fields are required to write the information into the MTJ cell. The fields are created by the superposition the magnetic field induced by current pulses in perpendicular lines according to the Stoner-Wohlfahrt principle. For this reason, the device was named field-induced magnetic switching (FIMS) approach or **Stoner-Wohlfahrt** approach (SW-MRAM). The scalability of devices based on this concepts turned out to be limited because of the required currents in the programming lines and the fact that the current densities in these lines exceeded physical limits below a certain feature size *F*. A solution to this problem has been offered by the spin-transfer phenomenon, theoretically predicted by Slonczewski and Berger in 1996 [12, 13] and experimentally confirmed few years later [14]. The **spin-transfer torque (STT)** phenomenon reflects the fact that a spin-polarized current flowing through a magnetic (nano-)structure can influence its magnetic state. This is due to the exchange interaction between the spin of the incoming conduction electrons and the spin of the electrons responsible for the local magnetization. The shrinkage of the cross section of the structures in the course of further miniaturization and the corresponding increase of the current densities is now an advantage since it enhances the STT effect. This effect is exploited in **STT-MRAM**s. Details are provided by Lucian Prejbeanu in Chapter D1 and, for example, in Ref. [15].

Electrostatic effects

A variety of electrostatic effects has been proposed in the literature to explain memristive phenomena. Some of them turned out to be misinterpretations of observations. Often, a trapping/detrapping of an electronic charge in the path of a conducting channel has been suggested. This is based on the idea that the (very small) leakage current in a capacitor with a highly insulating dielectrics can be modified by charges trapped in the dielectrics. While this is correct for an insulating dielectrics, the retention time of the element decreases with its resistance, similar to the RC time $\tau = RC$ of a capacitor (here, the capacitor is the memristive element). A simple estimate shows: In order to read the ON state of a highly scaled memristive element (e.g. an element with a cross section of 10 x 10 nm^2) in an array within less than, e.g., 100 ns, a current of at least 100 nA is required because the bitline capacitance C_{BL} of (at least) 30 fF must be charged to the read voltage (e.g. 0.3 V) in that time ($I = C\dot{V}$). This results in a current density of 10^5 A/cm^2 through the memristive element, i.e. the element must be quite conductive to carry this density. As shown in detail by Schroeder et al. [16], the retention time of the element will be less than the read time, which renders the device useless! Of course, this estimate shows an extreme case. For some applications such as neuromorphic circuits with relaxed requirements on small element sizes and high speeds, the situation may be somewhat different. However, in all conceivable realistic cases the retention times of memristive elements operating on trapping/detrapping of electron charge will be too low to be of practical use.

The only solution to this problem and, thus, the way to create a true **electrostatic trapping/detrapping element** is a completely different geometry beyond a simple MIM structure, i. e. introducing a lateral substructure into the I layer. One needs to move the trap site out of the flow of electrons. If there is a conducting channel, the trap site must be beside it. In order to charge and discharge the trap, another gate finger needs to be placed in the I layer. In principle, this is a floating gate MOSFET turned into a two-terminal device by connecting the source and the gate, and it will be introduced as a **MemFlash** in Chapter D9 by Martin Ziegler et al.. Possibly, the so-called **nanometal cell** concept nanosized dispersions of electronically conductive phases, e.g. metals such as Pt, in insulators such as SiO_2 offered by the group of I-Wei Chen [17, 18] follows the same principle, although there are still many open questions.

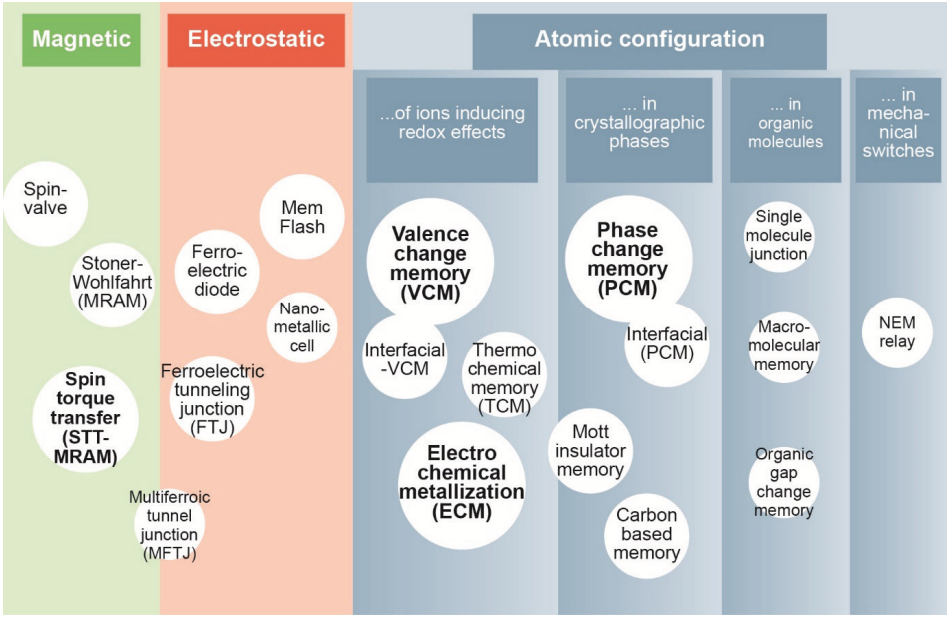


Figure 7: *Survey of memristive phenomena. Abbreviations and operating principles are explained in the text. Please note such a classification cannot be done without ambiguity. The type of phenomena and their weighing is qualitatively based on the number of publications, the degree of development, and the potential seen by the committee of the International Technology Roadmap for Semiconductors [32]*

Ferroelectric materials exhibit a spontaneous electric polarization which can be switched between two possible directions by an electrical field. Despite the fact, that the nature of ferroelectricity is a cooperative phenomenon that lies in a tiny displacement of ions in the crystal lattice, we will list the corresponding memristive effects as electrostatic effects. In general, the **ferroelectric memristive effect** is categorized into two major types depending on the conduction mechanism: a **ferroelectric tunneling junction (FTJ)** first proposed by Esaki in 1971 [19] and a **ferroelectric diode** first realized by Blom et al. in 1994 [20]. The ferroelectric tunneling junction consists of an ultrathin ferroelectric tunneling barrier. A reversal of its ferroelectric polarization is supposed to induce a change in the tunneling barrier height, giving rise to a change of the resistance of the cell. As a result of the domain-switching kinetics, different resistance levels can obtained by varying the electrical stimulus. The barrier-height modification due to polarization reversal requires an asymmetric potential distribution in the ferroelectric barrier as discussed by Tsymbal and Kohlstedt in 2006 [21]. There are several reports which indicate the experimental realization of this effect [22-24]. Great care must be taken in the interpretation, because the electric fields are so large that also other effects such as redox-based switching may be activated. Ferroelectric diodes are based on ferroelectric oxides which are turned into a semiconductor by extracting oxygen during processing. If such a semiconducting ferroelectrics is placed between a Schottky contact and an ohmic contact, the Schottky barrier is modulated by the direction of the ferroelectric polarization. Recent

reports are provided in Refs. [25-29]. A more comprehensive description of ferroelectric memristive devices is given in Chap. D9 and, e.g., in Chap. 16 of [4].

The idea of multiferroic memristive elements, in particular **multiferroic tunnel junctions (MFTJ)** combines the MTJ and the FTJ functionalities in one device. First studies have used combinations of ultrathin ferroelectrics such as $Pb(Ti,Zr)O_3$ and ferromagnetic oxides such $(La,Sr)MnO_3$ [30] or multiferroic oxides such as $BiMnO_3$ or $BiFeO_3$ have been employed [31].

The largest group of effects which lead to memristive phenomena are based on the **configuration of atoms** as the state variable. We will subdivide this group of effects into (a) those effects which rely on motion of ions on the nanoscale and related redox effects modifying the resistance of the material, (b) effects which are based on phase transitions, (c) effects which originate from configuration changes in organic molecules, and (d) effects of nanoelectromechanical switches.

Atomic configuration - of ions inducing redox effects

There is a broad range of ionic materials I in MIM cells in which the field-induced motion of ions on the nanoscale (*nanoionic* effects) leads to internal reduction-oxidation (redox) processes and, as a consequence, a change in the resistance of the cell. Three major types of such nanoionics redox-based memristive devices can be distinguished [33, 34]. In the engineering literature, these memristive devices are often denoted **redox-based resistance random access memories (ReRAM).**

Firstly, the **valence change memory** effect (**VCM**) occurs in a wide range of metal oxides and is (typically) triggered by a migration of anions, such as oxygen anions which are usually described by the motion of the corresponding vacancies, i. e. oxygen vacancies. Also cation interstitials may be the migrating species. Both, oxygen vacancies and cation interstitials, act as mobile donors. A subsequent change of the stoichiometry leads to a redox reaction expressed by a valence change of the cation sublattice and a change in the electronic conductivity in front of an electrode interface. This bipolar memory switching is induced by voltage pulses, where the polarity of the pulse determines the direction of the change, i.e. reduction or oxidation. In many cases, the switching occurs in conducting filaments in the VCM cell. In some cases, the switching takes place over the entire cross section of the electrode interface (**interfacial VCM**). Details will be given in Chap. D3 by Dittmann and in Chap. D4 by Menzel and Waser with respect to the kinetics. Some aspects of interfacial VCM cells are also covered in D9 by Ziegler and Kohlstedt. It is worth to mention, that there are also reports about a bulk-type memristive switching involving the entire I layer in MIM cells [35].

Secondly, the bipolar **electrochemical metallization** memory effect (**ECM**) which is also called Conductive Bridge RAM (CBRAM) relies on an electrochemically active electrode metal such as Ag, the drift of the highly mobile Ag^+ cations in the ion conducting I-layer, their discharge at the (inert) counter electrode leading to a growth of Ag dendrites which form a highly conductive filament in the ON state of the cell. Upon reversal of polarity of the applied voltage, an electrochemical dissolution of these filaments takes place, resetting the system into the OFF state. Details are described by Valov in Chap. D2 and, e.g., in Ref. [36].

A third type relies on a **thermochemical memory** effect (**TCM**, also called fuse-antifuse memory) due to a current-induced increase of the temperature which leads to a redox-related change of the stoichiometry along a discharge filament, and a subsequent freezing-in of this ON state. A differently shaped current pulse disrupts the conductive filament again to return the cell into the OFF state. Due to an inherently high energy consumption and strong variability of the parameters, R&D activities significantly decreased in the last five years. For details see Ref. [37].

From a historical point of view, this type of resistive switching has been studied in various solid-state materials since the early 1960s. Bistable resistance switching was reported in 1964 in NiO thin films on Ni substrate, where the switching was believed to be due to the formation and rupture of a nickel metallic filament in the NiO layer sandwiched by two electrodes [38]. Later in 1965, bistable resistive switching between two stable resistance states was shown in Nb_2O_5 [39]. Figure 8 shows the reported *I-V* curves Bi/Nb_2O_5(125 nm)/Nb, measured after forming (dielectric breakdown) of the initially insulating stack. Upon a first positive voltage sweep, the device is in a low resistance state (a). Application of a negative voltage leads to a RESET transition to high resistance (b), while application of a positive voltage causes the SET transition to the initial low-resistance state (c). The bipolar switching is bistable in that both states are stable.

Studies on bistable resistive switching have also been reported for thin films of Ta_2O_5 [40], SiO [41], TiO_2 [42], Al_2O_3 [43], and for ZnSe-Ge heterostructures [44]. The first report on what we today classify as ECM-type switching dates back to 1976 when Hirose and Hirose observed Ag dendrites being formed and dissolved between the Ag and Au electrodes in a bipolar operation mode of lateral Ag/As_2S_3/Au cells[45]. This early period of research faded in the late 1970s. Obviously the interest in this area decreased because of the overwhelming progress of the Si-based integrated circuit technology, in particular, the Flash memories. Another reason for the decrease in research in metal oxides and related compounds was presumably the lack of progress in understanding and controlling these resistive switching phenomena possibly due to insufficient analytical tools at that time. The period has been reviewed comprehensively by Dearnaley et al.[46], Oxley et al. [47], and Pagnia et al.[48].

A new era in research on resistive switching gradually started in the mid 1990s. The Tokura group found electrically triggered resistive switching in $Pr_xCa_{1-x}MO_3$ (PCMO) while investigating the magnetoresistive properties of this material [49]. In 2002, the IBM Zurich lab reported the resistive switching of perovskite-type zirconates, including many properties which are essential to NVM applications [50]. In the ECM-type area, Kozicki, Mitkova et al. started to study the Ag-GeSe systems in the late 1990s [51], while the Aono group published their first report on so-called atomic switches in 2001 [52]. These devices make it possible to control the electrochemical formation and dissolution of, for example, an Ag atomic bridge in a nanogap between a mixed electronic-ionic conducting Ag^+ electrolyte and a metal electrode with the precision of Landauer contance quantization [53]. In 2004, Samsung successfully demonstrated a high-density ReRAM chip using a 180 nm technology. It was based on unipolar switching Pt/NiO/Pt cells with an endurance of 10^6 SET/REST cycles. These and related papers have been the beginning of an unprecedented rise of R&D activities which led to the mega-trend which we encounter in the 2010s.

While there has been a basic understanding of the ECM mechanism from the beginning, a broad spectrum of mechanisms has been suggested as underlying mechanisms for the resistive switching in the various metal oxide systems. In 2005, Rainer Waser´s group was able to clarify the effect as a motion of oxygen ions and a coupled valence change in the cation sublattice on the nanometer scale at structural defects in the crystal lattice of the metal oxides near one electrode [54, 55]. For this reason, the expression valence change memory effect (VCM) was suggested for bipolar metal oxide systems [33], extending the more detailed classification of the nanoionically driven, redox process based resistive switching memories [56]. In 2008, Stan William´s group at the Hewlett-Packard Labs discovered that the electrical characteristics of bipolar resistive switching elements can be described in terms of the theory of memristive devices [57], and this link led to a further increase in the international research activities.

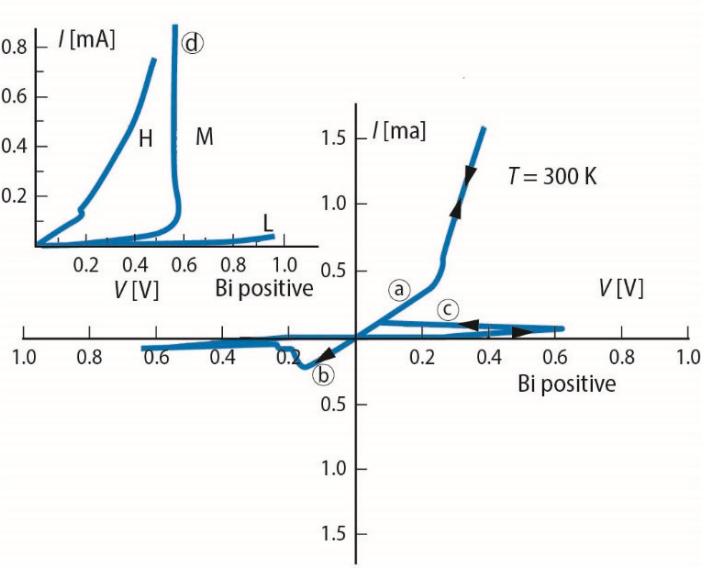


Figure 8: *Measured I-V characteristics for a Nb-Nb_2O_5-Bi MIM stack showing bistable resistive switching (redrawn after Ref. [39]). The device is initially in a low resistance state (a) due to the previous forming operation. Reset transition to the high resistance is shown for negative applied voltage (b), while set transition to the low resistance appears at positive voltage (c). The inset shows the I-V curves of three stable states, a high resistance state H and a low resistance state L and an intermediate state M (d).*

Atomic configuration – in crystallographic phases

In 1966, Ovshinsky described a unipolar, thermally driven switching between the amorphous and the crystalline phase of Ge-Te based compositions [58, 59]. In the 1980s phase change alloys on the pseudo-binary line between GeTe and Sb_2Te_3 with improved crystallization speeds were identified by Yamada et al. [60]. This discovery led to the successful development of optical storage based on phase change materials in the 1990s. Phase change re-writable optical storage uses the difference in reflectivity between the amorphous and crystalline phases to store information, and switching and reading is performed by laser pulses. **Phase change random access memory (PCM)** were developed in the mid-2000s and become one of the important emerging non-volatile memory technologies. They utilize the large difference in electrical resistivities between the two phases. Switching and reading is done using electrical pulses. To amorphize, a high laser/current pulse with short trailing edge is applied for melt-quenching. To crystallize, longer and lower intensity laser/current pulses are used, and even lower laser/current pulses measure the reflectivity/resistivity without causing any phase changes. Chalcogenide-based phase change materials are characterized by a unique property portfolio; they can be rapidly and reversibly switched between the amorphous and the crystalline state, which differ significantly in their properties. The group of Mattias Wuttig has identified the bonding mechanism responsible for these remarkable properties and has created a 'treasure' map, which reveals where such materials can be found and how their properties can be optimized [61]. Details

will be given in Chap. D6 by Matthias Wuttig emphasizing the materials aspects and in Chap. D7 by Martin Salinga with respect to devices and switching kinetics.

In order to decrease the switching energy in PCM cells, a new class of materials has recently been proposed by Simpson et al. [62] based on superlattice structures made of GeTe and Sb_2Te_3 layers. It has been suggested that the lower switching energy originates from the fact that the two relevant states are both crystalline and the transitions are constrained to atomic motion in one dimension. This new family of PCM materials has been called **interfacial PCMs (IPCMs)** and it is described in Chap. D8 by Riccardo Mazzarello.

For narrow band-gap Mott insulator compounds AM_4X_8 (A = Ga, Ge; M= V, Nb, Ta; X = S, Se), another unipolar memristive effect has been reported which is apparently related to a **Mott insulator-metal transition** triggered by an electronic avalanche effect above a critical field strength. The effect may be described as an electronic phase change effect in the crystalline phase in combination with local strain [63] and will be presented in Chap. D5 by Etienne Janod.

In MIM cells made from carbon (either carbon nanotubes layers or amorphous/nanocrystalline carbon thin films) unipolar memristive switching has been observed (**carbon-based memories**). Presumably it is induced by a thermal process which changes the amount of sp^2– and sp^3–hybridized C atoms along a filamentary region between the electrodes [64-65, 65]. Voltage pulses of moderate amplitudes and moderate rise/fall times are used for the SET operation and result in the formation of sp^2–rich, conductive regions, while short (ns) RESET pulses of larger amplitude lead to a disordered, sp^3–rich quenched state. The mechanism can be classified somewhere between a phase change mechanism and a thermochemical mechanism. In the latter, the change in the hybridization may be regarded as an intramolecular redox process in which the sp^2– and the sp^3–state represent the reduced and the oxidized C-atoms, respectively.

Atomic configuration – in organic molecules

There are two distinct areas in which the atomic configuration of organic molecules is proposed for memristive elements.

Single molecule junctions aim at individual contact to *single* molecules or small arrays of identical molecules arranged in a plane in a controlled manner. This approach tries to utilize the physical properties of single molecules for nanosized electronic devices. The invention and development of scanning probe techniques and many advances in micro- and nanotechnology have allowed the manipulation and operation of molecules on the level of small numbers or even individual objects. Still, by far most of the reports of single molecular switches (which occurred since the early 2000s) turned out to be due to preparation or measuring artefacts instead of being caused by true molecular features. An example, in which careful control experiments have been used to confirm that a configuration change of individual molecules in mechanical break junctions causes a memristive effect is reported in Ref. [66].

Bulk Molecular Systems for electronic devices are based on organic compounds with specific dielectric or electronic conduction properties. The organic compounds consist of small molecules, oligomers, or polymers and have found application in devices such as liquid crystal displays, organic light-emitting diode (OLED) displays, and soft organic transistors. The characteristic dimensions are much(!) larger than the sizes of the molecules. Consequently, most of the molecules are in (arbitrary) contact to other molecules, instead of being directly contacted by external electrodes. At the electrode interface, a huge ensemble of molecules is contacted by a typically inorganic, electronically conducting phase. At least two subgroups of bulk organic systems can be distinguished. MIM structures with polymers, in some case with nanodispersed metal particles, as insulators are often called **macromolecular memory** [32]. The memristive switching mechanism is not yet clear. And there is a class of bulk organic systems with similar

structures as OLEDs in which the band gap of the central layer can be modified by changing the atomic configuration of the photochromic molecules through an electric (or optical) stimulus [67]. In Fig. 7, we call this type **organic gap change memory**.

Another note of caution should be given. For several decades, there have been reports of *organic* materials which show resistive switching (see e.g. Ref. [68]). In some prominent cases it was discovered later that the switching, in fact, takes place in an oxide layer formed on an electrode metal used to contact the organic material. For example, Cu:TCNQ films sandwiched between Cu and Al electrodes [69] were found to switch resistively because of an ECM effect in the Al_2O_3 layer built during the processing of the system [70]. A similar situation was encountered for rose Bengal films between Al and Zn or ITO electrodes [71] for which the resistive switching presumably is caused by a VCM-type effect in the metal oxide layer of one of the electrodes [72].

Atomic configuration – in nanomechanical switches

The operation of electromechanical (EM) relays is based on the deflection of a flexible solid beam under the influence of an external force (of electromechanical, inverse piezoelectric, or electrostatic nature). Interestingly, the first computer in the modern sense, called Z3, was built in 1941 based on electromagnetic relays. Z3 consisted of 2600 relays, operated at 5 Hz, and used 60 V DC supply voltage. Nanofabrication technology today allow for the fabrication of nano-electromechanical relays, typically with electrostatic actuation, which can be used to build memristive switches (**NEM relays**). A survey is given in Ref. [73]. Because this area faded in recent years [32], these devices will not be covered in detail in this Spring School book.

3.4 Memristive Systems and Memristors

Throughout this Spring School book, we will use the term *memristive* in a qualitative manner denoting phenomena and devices in which the resistance state is memorized until the next stimulus which is strong enough to change the state (WRITE pulse).

In some journals it has become common to describe such non-volatile resistively switching devices in the framework of *memristors* defined by Chua in 1971 [74] and, more general, *memristive systems* introduced by him in 1976 [75]. In order to clarify some confusion encountered in the literature (to which some recent redefinitions contributed), we will briefly discuss the issue here. More details will be provided in Chapter E4. The *memristor* has been originally defined in 1971 by the relationship between the charge q and the flux φ (as the time integral of the voltage). In today´s terminology one would write:

$$\begin{aligned} V &= R(q)\cdot I \\ \dot{q} &= I \end{aligned} \tag{1}$$

This definition establishes the memristor as a fourth passive element apart from the resistor, the capacitor, and the inductor in the relation of the voltage V, the current I, the charge q, and the flux φ. However, until today, the memristor is a purely hypothetic element which is not represented by any simple device ("simple" means that electronic circuits which emulate its behaviour are excluded). Furthermore, even if it existed it would not be useful as a non-volatile resistively switching device because the time derivative of the charge q as the inner state variable is just proportional to I, while non-volatile resistively switching devices demand a very strong non-linearity (and other issues).

The general *memristive system* introduced in 1976 is a two-terminal device defined by a more complex state-dependent Ohm´s law and a state equation:

$$V(t) = R(\mathbf{x}, I, t) \cdot I(t)$$
$$\dot{\mathbf{x}} = f(\mathbf{x}, I, t) \tag{2}$$

$\boldsymbol{x}$ is an internal state variable or, more general, a vector of n internal state variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Internal state variables may be the temperature of the device, a magnetization, a chemical composition, etc. According to Eq. (2), the resistance R will be non-linear $R(I)$ and its value at any time t will depend on the entire past history of the device, because of

$$\mathbf{x}(t) = \int_{-\infty}^{t} f(\mathbf{x}, I, \tau) d\tau \tag{3}$$

.

In an I-V diagram, memristors show a hysteresis when V or I are used as a periodical stimulus. The hysteresis loop is *pinched*, i. e. it goes through the origin. The shape the hysteresis loop will depend on the frequency ω of the periodic stimulus. For $\omega \to \infty$ it will approach a linear resistance, and for $\omega \to 0$ it will approach a non-linear resistance. In this original definition of Leon Chua presented here, the memristive system represents, for instance, any type of thermistor (temperature-dependent resistor), or any other conceivable two terminal device. In order to use the concept of memristive systems to describe specifically non-volatile resistively switching devices, additional requirements must be introduced:

(1) x must include a *material dependent* state variable such as the magnetization, the crystallographic phase, or the length of a conducting filament formed by an internal redox process;

(2) x must have a lower and upper limit, $x_{min} < x < x_{max}$;

(3) since $\dot{\mathbf{x}}$ describes the kinetics of the switching process, the function f in $\dot{\mathbf{x}} = f(\mathbf{x}, I, t)$ must be highly non-linear in order to reflect a solution for the voltage-time dilemma.

4 Prospects and Challenges

Due to their attractive properties, several types of memristive devices such as STT-MRAM, ReRAM and PCM are promising for numerous applications, including memory and storage as well as digital and neuromorphic computing. Accordingly, these applications set a variety of requirements for the devices. The device performance requirements for memory and storage applications are more demanding than, for instance, for neuromorphic applications with respect to speed, retention time, and device density.

A range of memristive concepts offer a great scalability, ultra-fast switching speed, non-volatility, large HRS/LRS window, analogue resistance change, non-destructive reading, simple structures with common materials, 3D stackability, great CMOS compatibility and manufacturability. Apart from these great prospects, memristive device concepts often face a number of challenges. These challenges are application dependent. The most critical issues include device isolation in dense crossbar arrays, device variability, and reliability.

Based on progress in the microscopic understanding of the electroforming and switching process, solutions to the challenges will come from a combination of materials engineering, device structure optimization, as well as innovations in addressing/readout circuitry and programming algorithm.

Acknowledgments

The author would like to gratefully acknowledge fruitful discussions with many colleagues and financial support from the Deutsche Forschungsgemeinschaft through the SFB 917.

References

[1] R. Waser (Ed.). Nanoelectronics and Information Technology. Wiley-VCH (2012).

[2] R. Waser. Nanotechnology, Volume 3: Information Technology. Wiley-VCH, Weinheim (2008).

[3] An Chen, J. Hutchby, V. Zhirnov, and G. Bourianoff (eds). Emerging Nanoelectronic Devices. (2015).

[4] D. Ielmini and R. Waser. Resistive Switching - From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications. Wiley-VCH (2016).

[5] D. Kahng and S. M. Sze. A Floating Gate and Its Application to Memory Devices. Bell Systems Technical Journal, **46**, 1288-1295 (1967).

[6] R. Waser. Memoy Devices and Storage Systems - Introduction to Part V. Wiley-VCH, 603-620 (2012).

[7] E. Linn, R. Rosezin, C. Kügeler, and R. Waser. Complementary Resistive Switches for Passive Nanocrossbar Memories. Nature Materials, **9**, 403-406 (2010).

[8] S. Menzel, R. Dittmann, and R. Waser. Redox-based resistive switching - fundamentals and prospects. (2015).

[9] B. Dieny, V. S. Speriosu, S. S. P. Parkin, B. A. Gurney, D. R. Whilhoit, D. Mauri, and. Giant magnetoresistance in soft ferromagnetic multilayers. Phys. Rev. B, **43**, 1297 (1991).

[10] J. S. Moodera, L. R. Kinder, T. M. Wong, and R. Meservey. Large Magnetoresistance at Room-Temperature in Ferromagnetic Thin-Film Tunnel-Junctions. Physical review letters, **74**, 3273-3276 (1995).

[11] T. MIYAZAKI and N. TEZUKA. GIANT MAGNETIC TUNNELING EFFECT IN FE/AL2O3/FE JUNCTION. Journal of Magnetism and Magnetic Materials, **139**, L231-L234 (1995).

[12] J. Slonczewski. Current-driven excitation of magnetic multilayers. Journal of Magnetism and Magnetic Materials, **159**, L1-L7 (1996).

[13] L. Berger. Emission of spin waves by a magnetic multilayer traversed by a current. Physical Review B: Condensed Matter, **54**, 9353-9358 (1996).

[14] M. Tsoi, A. Jansen, J. Bass, W.-C. Chiang, M. Seck, V. Tsoi, and P. Wyder. Excitation of a magnetic multilayer by an electric current. Phys. Rev. Lett. **80**, 4281 (1998).

[15] B. Dieny, R. Sousa, J.-P. Nozieres, O. Redon, and I. L. Prejbeanu. Magnetic Randon Access Memories. Wiley-VCH, 655-668 (2012).

[16] H. Schroeder, V. V. Zhirnov, R. K. Cavin, and R. Waser. Voltage-time dilemma of pure electronic mechanisms in resistive switching memory cells. Journal of Applied Physics, **107**, 054517/1-8 (2010).

[17] A. B. K. Chen, B. J. Choi, X. Yang, and I. -W. Chen. A Parallel Circuit Model for Multi-State Resistive-Switching Random Access Memory. Advanced Functional Materials, **22**, 546-554 (2012).

[18] B. J. Choi, A. B. K. Chen, X. Yang, and I. Chen. Purely Electronic Switching with High Uniformity, Resistance Tunability, and Good Retention in Pt-Dispersed SiO2 Thin Films for ReRAM. Advanced Materials, **23**, 3847-3852 (2011).

[19] L. Esaki, R. B. Laibowitz, and P. J. Stiles. Polar Switch. IBM Tech. Discl. Bull. **13**, 2161 (1971).

[20] P. W. M. Blom, R. M. Wolf, J. F. M. Cillessen, and M. P. C. M. Krijn. Ferroelectric Schottky diode. Physical Review Letters, USA, **73**, 2107-10 (1994).

[21] E. Y. Tsymbal and Kohlstedt. Tunneling Across a Ferroelectric. Science, **313**, 181-183 (2006).

[22] V. Garcia, S. Fusil, K. Bouzehouane, S. Enouz-Vedrenne, N. D. Mathur, A. Barthelemy, and M. Bibes. Giant tunnel electroresistance for non-destructive readout of ferroelectric states. Nature, **460**, 81-84 (2009).

[23] A. Chanthbouala, A. Crassous, V. Garcia, K. Bouzehouane, S. Fusil, X. Moya, J. Allibe, B. Dlubak, J. Grollier, S. Xavier, C. Deranlot, A. Moshar, R. Proksch, N. D. Mathur, M. Bibes, and A. Barthélémy. Solid-state memories based on ferroelectric tunnel junctions. Nature Nanotechnology, **7**, 101 (2012).

[24] D. J. Kim, H. Lu, S. Ryu, C. W. Bark, C. B. Eom, E. Y. Tsymbal, and A. Gruverman. Ferroelectric tunnel memristor. Nano Letters, **12**, 5697 (2012).

[25] T. H. Kim, B. C. Jeon, T. Min, S. M. Yang, D. Lee, Y. S. Kim, S. H. Baek, W. Saenrang, C. B. Eom, T. K. Song, J. G. Yoon, and T. W. Noh. Continuous Control of Charge Transport in Bi-Deficient BiFeO3 Films Through Local Ferroelectric Switching. Adv. Funct. Mater. **22**, 4962 (2012).

[26] P. Maksymovych, S. Jesse,: P. Yu, R. Ramesh, and A.P. Baddorf: S.V. Kalinin. Polarization Control of Electron Tunneling into Ferroelectric Surfaces. Science, **324**, 1421-1425 (2009).

[27] T. Choi, S. Lee, Y. J. Choi, V. Kiryukhin, and S. W. Cheong. Switchable ferroelectric diode and photovoltaic effect in BiFeO3. Science, **324**, 63 (2009).

[28] An. Quan. Jiang, Can Wang, Kui Juan Jin, Xiao Bing Liu, James F. Scott, Cheol Seong Hwang, Ting Ao Tang, Hui Bin Lu, and Guo Zhen Yang. A Resistive Memory in Semiconducting BiFeO(3) Thin-Film Capacitors. Advanced Materials, **23**, 1277+ (2011).

[29] A. Tsurumaki-Fukuchi, H. Yamada, and A. Sawa. Resistive switching artificially induced in a dielectric/ferroelectric composite diode. Applied Physics Letters, **103** (2013).

[30] D. Pantel, S. Goetze, D. Hesse, and M. Alexe. Reversible electrical switching of spin polarization in multiferroic tunnel junctions. Nature Materials, **11**, 289 (2012).

[31] L. W. Martin, Y.-H. Chu, and R. Ramesh. Emerging Non-Volatile Memories. Springer, 103 (2014).

[32] The International Technology Roadmap for Semiconductors (ITRS). International Technology Roadmap for Semiconductors - 2013 Edition. (2013).

[33] R. Waser, R. Dittmann, G. Staikov, and K. Szot. Redox-Based Resistive Switching Memories - Nanoionic Mechanisms, Prospects, and Challenges. Advanced Materials, **21**, 2632-2663 (2009).

[34] R. Waser, R. Bruchhaus, and S. Menzel. Redox-based Resistive Switching Memories. Wiley-VCH, 683-710 (2012).

[35] Y. Aoki, C. Wiemann, V. Feyer, H.-S. Kim, C. M. Schneider, H. Ill-Yoo, and M. Martin. Bulk mixed ion electron conduction in amorphous gallium oxide causes memristive behaviour. Nature Materials, **5**, 3473/1-9 (2014).

[36] I. Valov. Redox-Based Resistive Switching Memories (ReRAMs): Electrochemical Systems at the Atomic Scale. ChemElectroChem, **1**, 26-36 (2014).

[37] D. Ielmini, R. Bruchhaus, and R. Waser. Thermochemical resistive switching: materials, mechanisms, and scaling projections. Phase Transitions, **84**, 570-602 (2011).

[38] J. F. Gibbons and W. E. Beadle. Switching properties of thin NiO films. Solid-State Electronics, **7**, 785-790 (1964).

[39] W. R. Hiatt and T. W. Hickmott. Bistable switching in niobium oxide diodes. Applied Physics Letters, USA, **6**, 106-108 (1965).

[40] K. L. Chopra. Avalanche-induced negative resistance in thin oxide films. Journal of Applied Physics, USA, **36**, 184-187 (1965).

[41] J. G. Simmons and R. R. Verderber. New thin-film resistive memory. Radio and Electronic Engineer, UK, **34**, 81-89 (1967).

[42] F. Argall. Switching phenomena in titanium oxide thin films. Solid-State Electronics, **11**, 535-541 (1968).

[43] T.W. Hickmott. Electroluminescence, Bistable Switching, and Dielectric Breakdown of Nb2O5 Diodes. Journal of Vacuum Science and Technology, **6**, 828-833 (1969).

[44] H. J. Hovel and J. J. Urgell. Switching and memory characteristics of ZnSe - Ge heterojunctions. Selected papers in high-polymer physics, Journal of Applied Physics, **42**, 5076-83 (1971).

[45] Y. Hirose and H. Hirose. Polarity-dependent memory switching and behaviour of Ag dendrite in Ag-photodoped amorphous As_2S_3 films. Journal of Applied Physics, USA, **47**, 2767-72 (1976).

[46] G. Dearnaley, A. M. Stoneham, and D. V. Morgan. Electrical phenomena in amorphous oxide films. Reports on Progress in Physics, **33**, 1129-1191 (1970).

[47] D. P. Oxley. Electroforming, switching and memory effects in oxide thin films. Electrocomponent Science and Technology, UK, **3**, 217-24 (1977).

[48] H. Pagnia and N. Sotnik. Bistable switching in electroformed metal-insulator-metal devices. Physica Status Solidi A, East Germany, **108**, 11-65 (1988).

[49] A. Asamitsu, Y. Tomioka, H. Kuwahara, and Y. Tokura. Current switching of resistive states in magnetoresistive manganites. Nature, **388**, 50-2 (1997).

[50] A. Beck, J. G. Bednorz, C. Gerber, C. Rossel, and D. Widmer. Reproducible switching effect in thin oxide films for memory applications. Applied Physics Letters, **77**, 139-41 (2000).

[51] M. N. Kozicki, M. Yun, L. Hilt, and A. Singh. Applications of programmable resistance changes in metal-doped chalcogenides. Proceedings of the Internat. Solid-State Ionic Devices Conf. Seattle, WA, USA, 02/05/1999-07/05/1999, Electrochem. Soc, 298-309 (1999).

[52] K. Terabe, T. Hasegaw, T. Nakayama, and M. Aono. Quantum point contact switch realized by solid electrochemical reaction. Riken Rev, **37**, 7-8 (2001).

[53] K. Terabe, T. Hasegawa, T. Nakayama, and M. Aono. Quantized conductance atomic switch. Nature, **433**, 47-50 (2005).

[54] R. Waser, K. Szot, W. Speier, R. Oligschlaeger, and S. Karthäuser. Resistive switching in oxide systems (inv. talk). 10 years of Quantum Science Research; HP laboratories, Palo Alto, March 25-27 (2005).

[55] K. Szot, W. Speier, G. Bihlmayer, and R. Waser. Switching the electrical resistance of individual dislocations in single-crystalline $SrTiO_3$. Nature Materials, **5**, 312-320 (2006).

[56] R. Waser and M. Aono. Nanoionics-based resistive switching memories. Nature Materials, **6**, 833-840 (2007).

[57] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams. The missing memristor found. Nature, **453**, 80-83 (2008).

[58] S. R. Ovshinsky. Symmetrical current controlling device. US patent, 3271591 (1966).

[59] S. R. Ovshinsky. Reversible electrical switching phenomena in disordered structures. Physical Review Letters, USA, **21**, 1450-3 (1968).

[60] N. Yamada, E. Ohno, N. Akahira, K. Nishiuchi, and K. Nagata. Overwritable Phase-Change Optical Disk Material. Japanese Journal of Applied Physics Part 1-Regular Papers Short Notes & Review Papers, **26**, 61-66 (1987).

[61] D. Lencer, M. Salinga, B. Grabowski, T. Hickel, J. Neugebauer, and M. Wuttig. A map for phase-change materials. Nature Materials, **7**, 972-977 (2008).

[62] R. E. Simpson, P. Fons, A. V. Kolobov, T. Fukaya, M. Krbal, T. Yagi, and J. Tominaga. Interfacial phase-change memory. Nature Nanotechnology, **6**, 501 (2011).

[63] P. Stoliar, M. Rozenberg, E. Janod, B. Corraze, J. Tranchant, and L. Cario. Nonthermal and purely electronic resistive switching in a Mott memory. Physical Review B: Condensed Matter, **90**, 45146/1- (2014).

[64] F. Kreupl, R. Bruchhaus, P. Majewski, J. B. Philipp, R. Symanczyk, T. Happ, C. Arndt, M. Vogt, R. Zimmermann, A. Buerke, A. P. Graham, and M. Kund. Carbon-based resistive memory. 2008 international electron devices meeting - technical digest, 521-524 (2008).

[65] A. Sebastian, A. Pauza, C. Rossel, R.M. Shelby, A.F. Rodriguez, H. Pozidis, and E. Eleftheriou. Resistance switching at the nanometre scale in amorphous carbon. New Journal of Physics, **13**, 013020 (2011).

[66] E. Loertscher, J. W. Ciszek, J. Tour, and H. Riel. Reversible and Controllable Switching of a Single-Molecule Junction. Small, **2**, 973-977 (2006).

[67] R. C. Shallcross, P. Zacharias, A. Koehnen, P. O. Koerner, E. Maibach, and K. Meerholz. Photochromic Transduction Layers in Organic Memory Elements. Advanced Materials, **25**, 469 (2013).

[68] T. Lee and Y. Chen. Organic resistive nonvolatile memory materials. MRS Bulletin, **37**, 144-149 (2012).

[69] R. S. Potember, T. O. Poehler, and D. O. Cowan. Electrical switching and memory phenomena in Cu-TCNQ thin films. Applied Physics Letters, USA, **34**, 405-7 (1979).

[70] T. Kever, U. Boettger, C. Schindler, and R. Waser. On the origin of bistable resistive switching in metal organic charge transfer complex memory cells. Applied Physics Letters, **91**, 083506-1-3 (2007).

[71] A. Bandyopadhyay and A. J. Pal. Multilevel Conductivity and Conductance Switching in Supramolecular Structures of an Organic Molecule. Applied Physics Letters, **84**, 999-1001 (2004).

[72] S. Karthaeuser, B. Lussem, M. Weides, M. Alba, A. Besmehn, R. Oligschlaeger, and R. Waser. Resistive switching of rose bengal devices: a molecular effect?. Journal of Applied Physics, USA, **100**, 94504-1-6 (2006).

[73] K. Akarvardar and H.-S. Philip Wong. Nanomechnical Logic Gates. Wiley-VCH (2012).

[74] L.O. Chua. Memristor-the missing circuit element. IEEE Transactions on Circuit Theory, **CT-18**, 507-519 (1971).

[75] L.O. Chua and S.M. Kang. Memristive devices and systems. Proceedings of the IEEE, **64**, 209-223 (1976).

A 1 Structure of Matter – From Perfect Crystals to Amorphous Materials

David P. DiVincenzo
Peter Grünberg Institut, PGI-2
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Crystals – the packing problem **2**

3 Ionically Bonded Crystals **5**

4 Covalently-bonded Crystals **7**

5 Defects in Crystals **8**
5.1 Planar Defects 8
5.2 Line defects: dislocations 10
5.3 The all-over defect: amorphous solids 11

6 Finally **13**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

You have known since early in your schooling that solid things are made of atoms, and that they become solid because of interactions, having at least some attractive component [1], between those atoms. "Condensed matter" comes in an incredible variety of forms; to contemplate all its manifestations requires turning one's mind to a wide flung set of disciplines, from biology to neutron star theory, from cosmology to earth science – and, of course, within the more familiar precincts of chemistry and physics. It is in these latter fields that the condensed states that are of interest for this School, and for this Chapter, will lie.

This chapter will provide an introductory basis for the interesting nano-phenomena that are at center stage in this School, by describing the basic solid structures that form the basis for these phenomena. We will begin with the various perfect crystalline solids, giving their basic structure, along with the reasons why they form as they do. Then we will move on to various of the ways that crystals can be imperfect, including the complete loss of order, but not of solidity, that is embodied in the amorphous state.

2 Crystals – the packing problem

While it will not be the main focus here, it is good to start with the simplest crystallization scenario, embodied by the packing of identical hard spheres. This is a suitable mathematical idealization of the case where atoms interact my simple two body force laws. This happens in two distinct situations: 1) for simple metals, where we have the "metallic bond", 2) for rare gasses, where the (weak) interaction is via the "van der Waals bond". The scheme for getting to the lowest energy state is simple: pack in the most neighboring spheres possible.

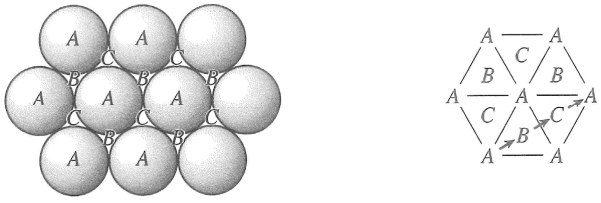


Fig. 1: *Dense packing of identical spheres. The ABCABC... sequence gives face-centered-cubic packing, but infinitely many other packings are possible. See [2] for many more details.*

Figure 1 illustrates the fruitstand-packing that I am sure that you all know, and the exact basis of many real crystalline structures. But a basic mathematical fact, which has myriad implications for the crystal structures that are based on this packing, is that it is not unique: there are infinitely many densest packings of identical spheres in three dimensions. The non-uniqueness has to do with the stacking of the triangle-packing 2D layers. Having placed a first layer, labelled "A", there are two possible positionings, "B" and "C", for the layer on top. For the second layer these are not really distinct, as one can always shift or rotate an "AC" bilayer so that is the "AB" arrangement. But for the *third* layer, no such equivalence exists: "ABC" and "ABA" packings are not geometrically equivalent. There are then infinitely many ways to continue – any letter sequence (no repetitions) like ABACBCBABC... is a valid closest packing.

Seen very commonly are the two simplest repetition patterns: ABCABCABC... and ABABABABAB... These are really different – the *second* neighbor environment of an atom is

different in the two cases, so there are small differences in energy of the two configurations. They also have completely different structural symmetries. The ABC case gives the face centered cubit packing, with a one-atom repeat unit (the right side of Fig. 1 shows why it is only one), while ABAB is called hexagonal close packing, with a two-atom repeat unit. The closeness in energy of these two different structures has many consequences for defect structures in these and related crystals, as will be seen below.

In the following section we will move on to similar mathematical sphere-packing considerations for compound solids, in which assemblies of different-sized spheres are the appropriate model. But first, I want to take a detour into a basic question which is implied by what I have just said: why are simple repetitive patterns like ABCABC... and ABABAB... preferred over more complex or random arrangements? I want to delve into this question by introducing, and commenting on, an extensive quotation from the work of P. W. Anderson [3]. He writes,

The essential phenomenon in either case is that the lowed state of potential energy of interaction between particles – for example, a pair interaction

$$V_{tot} = \sum_{ij} V\left(|r_i - r_j|\right)$$

– must occur for either a unique relative configuration of all the particles

$$C = \{r_1, r_2 \ldots r_N\}$$

(and all translations and rotations)

or, in artificial cases, perhaps for a highly restricted subset ... So far as I know, there exists no proof that among the lowest energy configurations C at least one is a regular lattice, but I for one would be very surprised if this weren't so.

Note that Anderson is claiming, in this rather informal statement, that periodic crystalline arrangements are always the lowest energy state. There immediately follows his argument for this:

One may work up a reasonable argument for it as follows: Let us take a relatively small box containing n atoms and consider its optimum configuration. There will be one minimum-energy configuration, all small displacements from which are described by a harmonic potential:

$$V - V_{min} = \frac{1}{2}\sum_{ij} V_{ij}^{"}\delta\, r_i \delta\, r_j$$

(incidentally, we may allow small changes in the shape of our box to minimize the energy further). The effect of the smallness of the box may be minimized by using periodic boundary conditions.

For large displacements, however, there may be additional relative minima: for instance, but not typically, a single atom may have two possible potential wells within which it might sit, one lower than the other... Such a second minimum will have an energy only of order unity above the true minimum; but in general the other configurations will typically have energies ≈ n higher (as in the regular case, lattice energy differences for different lattices are proportional to the size of the crystal; even for an irregular array moving every atom in an essential way will change the energy by ≈ n).

Now we imagine putting a much larger array together of N = mn atoms. An attempt at a low-energy configuration may be made by simple piling all the boxes of n atoms together and removing the interior walls. We can expect that small harmonic readjustments will further improve the energy; but our basic argument is that the energy of misfit between the surface of the small pieces is a surface energy, of order $n^{2/3}$, while the energy necessary to change the interior configuration in an essential way will be ≈ n, so that there will be a cell size, n, beyond which it will not pay to modify the internal configurations; this then gives us a regular array.

The final sentences are the crucial conclusions of this argument: repetitive arrangements of atoms are likely to be the lowest energy configuration. Note that Anderson indicates that his statements constitute a "reasonable argument" for crystallinity – he does not venture to use the word "proof". One might note that his arguments have greater force for large *n*, so that it would be an indication that crystals with large repeat units would be energetically favorable, and it really doesn't have much to say about simple crystals where *n* is 1 or 2.

It is well that he did not try to work up these arguments further into a mathematical proof, because the basic premise of his discussion was in fact proved *wrong*, in a very interesting way, in experiments in the 1980s. This was in the discovery of quasicrystals [5], very regular packings of atoms in which the lattice arrangement is not periodic! Figure 2 shows packing arrangements [4] seen in AlMn and AlFe compounds that are compatible with packings of certain rhombohedral shapes. These shapes do not fill space periodically, but they can fill space quasiperiodically. Quasiperiodicity has a particular technical meaning, in which the diffraction pattern of the solid shows only sharp spots, indicating that the material is not disordered in the usual sense.

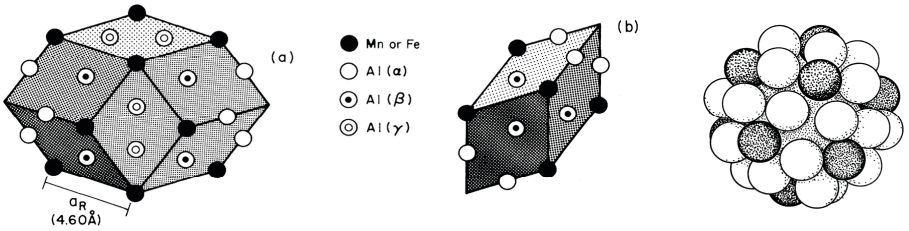


Fig. 2: *Local packing arrangements as observed in Al-Mn alloys, organized in the left two structures to show their relations to 3D tile shapes. These tiles, with the matchings implied by the atom positions on their faces, cannot tile space periodically, but can tile quasiperiodically. On the right is the 54-atom "MacKay icosahedron", another fragment of the same packing from [4].*

The fact that the packing of certain shapes enforces nonperiodicity was an insight of Penrose [6] in the 1970s (Fig. 3), although the original insight is apparently due to Islamic mosaic artisans of a millenium ago, as recounted in Penrose's paper. These quasicrystalline materials turn out to be quite varied, but none of them play a role presently in nanoswitch devices. So, for the rest of this chapter we will pretend that Anderson was right, and we will confine ourselves to solids for which the perfect system is crystalline, and imperfections are particular kinds of departures from crystallinity.

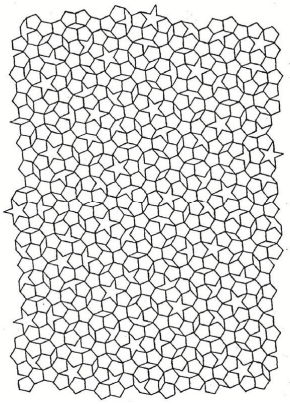


Fig. 3: *R. Penrose anticipated the discovery of alloy quasicrystals in the 1970s [6], when he observed that certain assemblies of tiles, with matching rules, permit only quasiperiodic space filling. This tiling is apparently the first of the many that he discovered.*

3 Ionically Bonded Crystals

We will now proceed to cases of compound solids, in which the cohesion of the crystal is understood to be due to the fact that these is charge transfer from one species, the "cation" to the other, the "anion" [7]. There are certain systematic tendencies that connect a few facts about the species involved and the crystal structure that results. Figure 4 shows the most common cases that arise for the simple diatomic ionic solids. On the left is the unit cell of the crystal structure of CsCl. You will recall from school that we consider Cs here to be a cation with charge +1, and Cl to be anionic with charge -1. Furthermore, it is deduced from their various appearances in the solid state that the radii of these two ions is almost equal. It is seen that many diatomic ionic compounds with equal ionic radii adopt this particular ionic packing.

The preferred sphere packing is observed to change as the ratio of anionic to cationic radii changes. The middle sketch shows the familiar case of salt, which contains the smaller cation Na^{+1}. We say that because of its smallness, there is only room to fit a smaller number of anion neighbors around it – in fact 6 in the form of an octahedron, instead of the 8 in a cube arrangement as in CsCl. Finally, there are cases of extreme contrast in ionic radii, for example the small Zn^{+2} ion with the large S^{-2} ion, for which the preferred arrangement is the zincblende (the word simply means "zinc sulfide") crystal structure on the right of Fig. 4. Here the small cation is surrounded by only 4 anions in a tetrahedral arrangement.

The reader will notice that I have used some rather indefinite words to indicate the connection between the ion properties and the crystal structure. The concept of ionic state is an immensely useful one in chemistry, but its literal reality is open to question. Many believe that it would be impossible, even in principle, to prove, using a minute electrometer probing the inner spaces of a

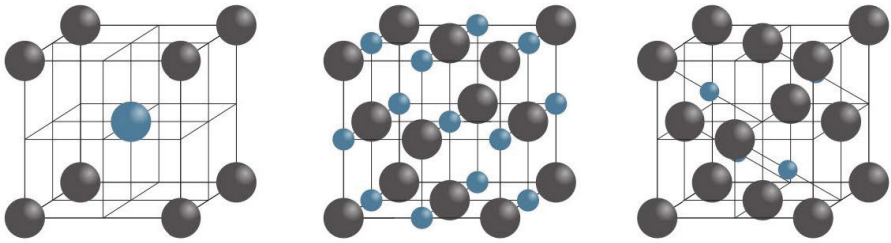


Fig. 4: *Ionic binary-compound crystal structures. Left: CsCl structure. Middle: NaCl structure. Right: ZnS structure. See [2].*

crystal, that the zinc atom really has a deficit of two electrons, as it could unambiguously have in an atomic experiment in which two electrons are ejected with ultraviolet light from a single zinc atom that is held in a trap. We will see in a short while that zincblende is a particular borderline case, in which we give a frankly schizophrenic view of the nature of the crystalline bonding. But the fact is that the concept of the solid-state ion is indispensible as an organizing principle for a vast set of crystalline materials.

In particular, the ionic-bonding concept is indispensible for explaining the atomic arrangement in some of the most important compound materials that appear in this school. Figure 5 shows the crystal structure of perovskite, which has the generic chemical formula ABX_3. The anion is almost invariably O^{2-} (there is a distinct class of compounds in which X=Ni), A is the large cation with formal charge +2, and B is the small cation with formal charge +4. $BaTiO_3$, $SrZrO_3$ and $CaSnO_3$ are important examples of these perovskites. Formal AB charges of +1/+5, and +3/+3 (e.g., $LaMnO_3$) also occur. Note that the designations "small" and "large" are justified by the differing number of anion nearest-neighbors of the B and A species: 8 (octahedral) for the B, 12 (cuboctahedral) for the A. One can focus on the oxygen octahedra with the small cation inside, and consider the perovskite lattice as a regular network of vertex-sharing octahedra. For many cases, the resulting crystal retains the full symmetry of the octahedron. However, outside the "preferred" range of cationic radii, the ground state of the crystal prefers some distortion (rotation, tilting) of the linkages between these oxygen octahedra, leading to crystals with lower symmetry, for example trigonal [8].

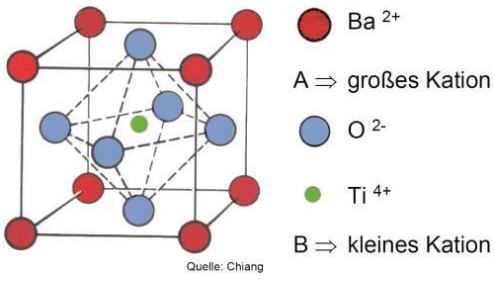


Fig. 5: *The perovskite crystal structure.*

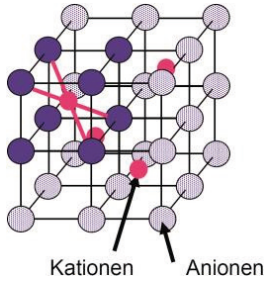


Fig. 6: *The crystal structure of fluorite.*

While the variety of ionic oxide crystal structures is vast, I will content myself with mentioning just one of the most important additional types, that represented by fluorite (CaF_2). This AX_2 structure is shown in Fig. 6. Many important crystals come in this form, including those of ZrO_2 and CeO_2. In fact, the list of fluorite compounds is vast, here is an interesting list: $PtGa_2$, $SnMg_2$, $AuIn_2$, TbO_2, UO_2, EuF_2, HgF_2, CeH_2, LuH_2, NbH_2, UN_2.

4 Covalently-bonded Crystals

Now we turn to those compounds in which the rationalization of the structure comes from reasoning concerning chemical bonding, rather then the efficient packing of charged ions. Chemical bonding introduces directionality constraints that result sometimes in "looser" packings than one would arrive at by simple sphere-packing considerations.

We begin with the familiar "diamond" lattice, see Fig. 7. We have shown it in two-sublattice form; for elemental C, Si, and Ge both sublattices are identically occupied, but for compound structures that we will mention shortly they are occupied by two different species. Each sublattice is a face-centered cubic lattice, and is thus a densest-sphere packing. For the most part this has only formal significance, although it gives us some explanatory power when we come to the discussion of planar defects later in this chapter.

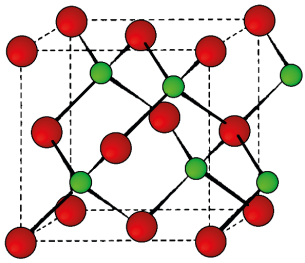


Fig. 7: *The zincblende (again) or diamond crystal structure, emphasizing its covalent character.*

Why do the group IV elements crystallize in this "loose" manner? The chemical explanation is as follows: the s and p electronic orbitals containing the four valence electrons of these elements undergo a hybridization, forming sp_3 linear combinations. The sp_3 orbitals are highly directional, and they are pointed in the four tetrahedral directions. Two such orbitals from two neighboring atoms strongly overlap, and the two electrons contributed from each atom form the bonding pair. Thus, the tetrahedral nearest neighbor arrangement of the crystal of Fig. 7 can optimally take advantage of this tetrahedral bonding geometry.

Almost all the AB compounds of group III and group IV elements assume exactly the same crystal structure. The chemical explanation is that A^- and B^+ are isoelectronic to a group-IV element, and then the story of tetrahedral sp_3 direction, perhaps supplemented by some ionic adhesion, explains the stability of these crystals.

The crystal structure of Fig. 7 is also that of some II-VI AB compounds, and one is tempted to consider A^{-2} and B^{+2} to be isoelectronic again to group IVs and use the sp_3 explanation. Recall, however, that the crystallization of the II-VI zinc sulfide has already been "explained" by ionic packing considerations – note in fact that the zincblende crystal structure is identical to the "diamond" crystal structure of Fig. 7! This illustrates that the chemical reasoning that explains these crystal forms is by no means rigorous. The physicist, who uses a large numerical calculation using, e.g., density functional theory to expain the stability of these crystals does not "need" sp_3

or ionicity ideas. Nevertheless, when used wisely, the explanatory power of the chemical reasoning is very great, and is not to be discounted.

The strong tendency of Si to form tetrahedral bonds is also manifest in another important material, SiO_2. Fig. 8 shows the crystal structure of quartz, one of the important crystalline modifications of silicon dioxide. The tetrahedrally coordinated oxygens surrounding each silicon form a network of vertex-sharing tetrahedra. The vertex connection is "bent", and this has another chemical explanation: the group-VI oxygen has two unpaired electrons available for bonding, and the other two are non-bonding "lone pairs". These two lone pairs, and the two resulting bonding pairs, form nearly a tetrahedron around the oxygen, and the bonds have nearly the tetrahedral angle. (The same explanation is given for the "water angle" in H_2O.) This explanation should again be put in the category of very non-rigorous, but extremely useful. But Fig. 8 shows another experimental fact about this oxygen bonding: the "thermal ellipsoids" of the atoms in the crystal are shown, that is, the extent of motion of the atoms due to thermal excitations at normal temperature. The oxygen bond angle is evidently quite "floppy" compared with that at the silicons. This floppiness will come back again, when we considered disordered forms of covalent crystals in Sec. 5.3.

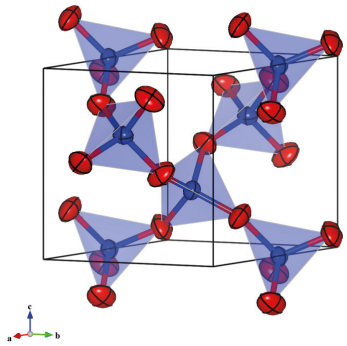


Fig. 8: *The crystal structure of SiO_2, quartz. The ellipsoids show the degree of thermal motion of the oxygen (red) and silicon (blue).*

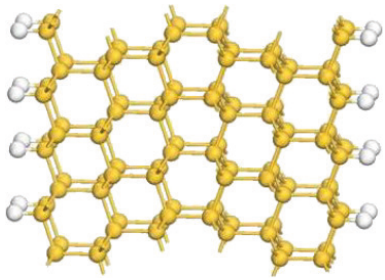


Fig. 9: *Twinning in the diamond crystal.*

5 Defects in Crystals

5.1 Planar Defects

There is a vast variety of ways in which covalent or ionic solids can depart from regularity. Here we will survey just a few interesting ways in which this irregularity manifests itself. We begin with defects defined by planes. The next section discusses defects as lines; defects as points will be covered within lecture A3 in this school.

Recall the mathematical fact that irregularities can readily occur in the stacking of close-packed planes of spheres. Since the diamond lattice is, mathematically, two interpenetrating fcc lattices (the AB sphere-packing lattice), it is also subject to this stacking variation. Figure 9 shows such a "stacking fault". Chemical reasoning based on sp_3 orbital chemistry would deem this

structure to be isoenergetic to the perfect diamond lattice: all bond lengths can be unchanged, and all bond angles can remain exactly tetrahedral. Of course, this stacking fault *does* cost a finite energy; the geometry of second-neighbor atoms is altered near the stacking fault, so one can say that there are small longer-distance interactions that are not optimized here. Connected with this, there are changes of the bond lengths and bond angles in the vicinity of the stacking fault, but again these changes are very small.

As Fig. 10 illustrates, many variants on stacking faults, connected with the variations of the ABC stacking noted in Sec. 2, also readily occur in crystals of elements in group IV.

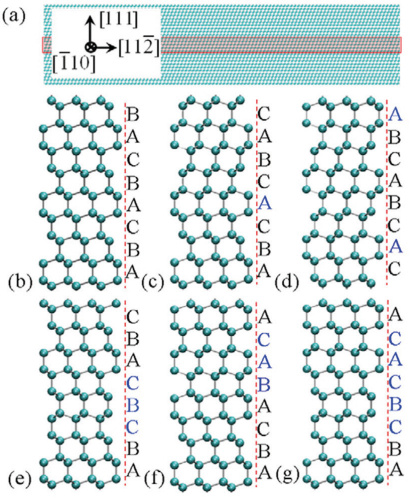


Fig. 10: *Faulted stacking arrangements as they are observed to occur in Si nanowires. See [9].*

Returning to Fig. 9, one can see that besides being viewed as a change of stacking, this planar defect can be viewed as two half-spaces of perfect crystal, differently oriented in space, and jointed together at the plane. From this point of view, this is a very special example of the general category, the *grain boundary*. This is an especially symmetric grain boundary with mirror symmetry across the plane, making this a *twin boundary*, or *twin* for short. The special orientation angle consistent with the alternate stacking described above makes this a very common twin, and it is called a *primary twin* for these crystals (it has the designation Σ3 in crystallography, but this is not so important for us here.

Figure 11a shows a simple mechanism by which twins of different orientations can arise [10]. If primary twinning occurs on two equivalent but different crystallographic planes (i.e., the meeting planes of the A and B domains and of the A and C domains in Fig. 11a), then in the course of further growth the domains B and C should join at a common plane. If the growth occurs in the most symmetrical way as shown, then another twin, the so-called Σ9 second-order twin, is formed. Its atomic structure, as shown in Fig. 11b [10], involves some clear departures from the near-perfection of the stacking fault: cond angles and lengths are distorted, and the normal six-fold ring structure is replaced by five- and seven-fold rings in the grain boundary plane.

5.2 Line defects: dislocations

We have just seen an example of a line defect, the tricrystal line in Fig. 11. But the more usual line defect that we consider is the *dislocation*. Ref. [11] has discussions of most of the points that I will make in this section. Figure 12 shows, for an abstract cubic lattice, one type of dislocation, the so-called screw dislocation. This can be thought of as a structure that results if the crystal is cut along a half-plane, and then "glued" together with some shifting of the lattice. This shift is called the "Burgers vector". This vector is obtained by taking a circuit (the so-called "Burgers circuit") on the crystal lattice along a path enclosing the line defect. One chooses a path (e.g., 2 steps north, 4 steps west, ...) that would close if the defect were not present. The presence of the dislocation defect causes a "closure failure", which is the difference between the points E and F in Fig. 12. The connecting vector between these two points is the Burgers vector. The screw dislocation is characterized by the Burgers vector being parallel to the line defect.

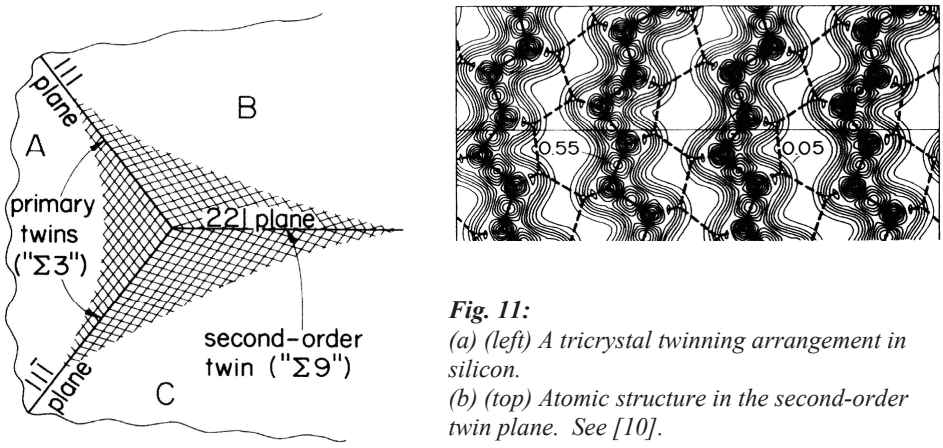


Fig. 11:
(a) (left) A tricrystal twinning arrangement in silicon.
(b) (top) Atomic structure in the second-order twin plane. See [10].

In the other important dislocation type, the edge dislocation, the Burgers vector is perpendicular to the line of the defect. See Fig. 13 for the abstract depiction of this defect; note the closure failure on the ABCD circuit shown.

Note that both Figs. 12 and 13 are highly unrealistic as atomic structures – note the very lattice points that result from the naive "gluing". However, these problems occur only near the "core" of the defect, and the slightly distorted crystal that results a few lattice constants away from the core is quite realistic. The elastic distortion field is in fact well represented by considering the Burgers vector as a "pole" source, very analogous to a line source of electric dipoles in electrostatics. In the core, it is mandatory that there be some modified atomic arrangement, whose details are specific to the crystal type. But dislocations are found in all types of crystals, no matter what the type of atomic interaction – van der Waals, metallic, ionic, or covalent.

The previous discussion would suggest that the Burgers vector must be a full lattice translation of the crystal lattice. But in crystals like diamond in which there is more than one atom per crystal unit cell, there is another possibility: one also readily observes *partial dislocations*, in which the Burgers vector is smaller, typically one half of a full lattice translation. For such a dislocation, there is not perfect, elastically distorted crystal in all directions away from the core;

in fact, the core is then the starting line of a stacking fault, which ends (often quite nearby) at another partial dislocation. One can in these situations consider the partials to be the result of the dissociation of a full dislocation into two equivalent parts.

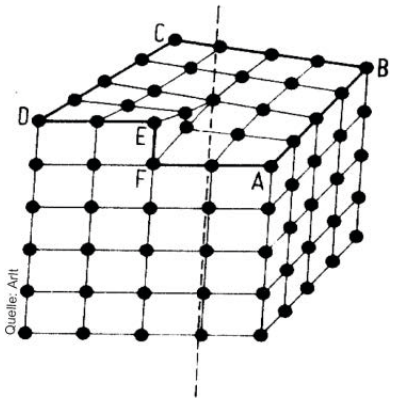


Fig. 12: *The screw dislocation.*

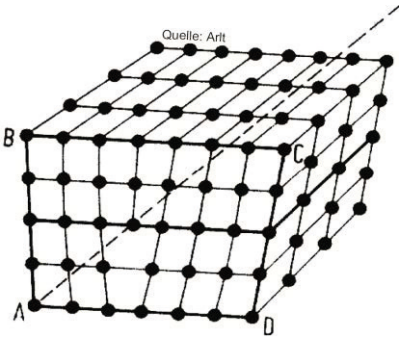


Fig. 13: *The edge dislocation.*

The mobility of dislocations is another matter, whether dislocations can move under the influence of stresses is another matter. Dislocations tend to be most mobile in crystals with non-directional bonding; in metals dislocations move readily, and the plastic deformation that occurs when you bend a paperclip results from the motion of "billions and billions" of dislocation lines. A most common motion mechanism is the glide motion of an edge dislocation, in which also can be seen the connection between dislocation and deformation of the crystal. In Fig. 14, the glide process is visualized, likened to the locomotion of a caterpillar. The important point about the glide process is that, by very small motions of individual atoms, a whole half-plane of atoms (red) end up transported from one side of the crystal to the other; thus plastic deformation is associated with this "slip plane" (along which many glide events occur). Pileups of these dislocations are also associated with the fracture of crystalline material.

5.3 The all-over defect: amorphous solids

The final case that we will touch in this lecture contains solids that are so disrupted in their atomic order that no crystal lattice is left: the solid is said to be amorphous, but that is not to say that it has lost all vestige of structure. In fact, typically, amorphous solids have a large degree of local ordering that is very reminiscent of the structure found in related crystals.

There are many elements that have solid non-crystalline or amorphous forms, but to this writer's knowledge all examples of such arise in the covalently bonded elements. Elements with non-directional bonding (van der Waals or metallic) seem to always manage to crystallize, perhaps with a rather dense network of grain boundaries (the exception that proves this rule is apparently Xe [12]). Atoms in this case apparently find it very easy to slip into their places in a dense-packing sphere arrangement.

All of the group IV elements can form amorphous solids. The "looseness" of the tetrahedral coordination that is preferred by these elements is quite compatible with random space filling,

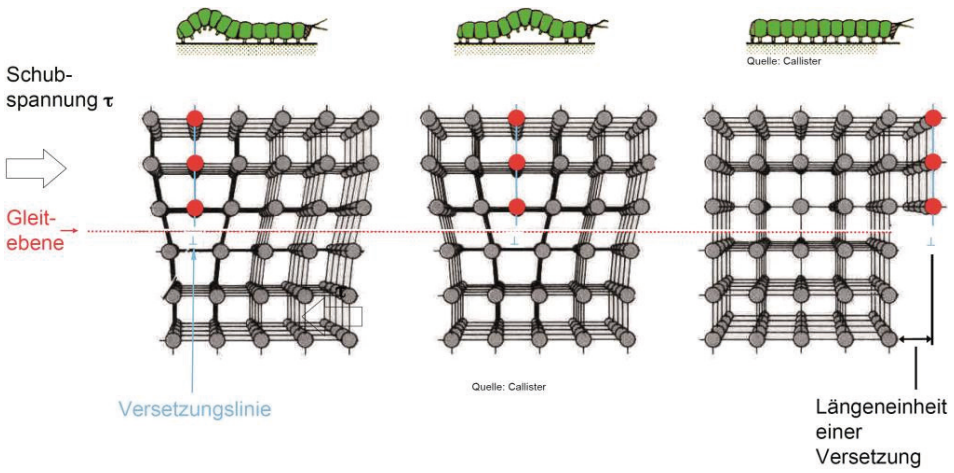


Fig. 14: *The glide motion of an edge dislocation on its slip plane.*

as can be seen in Fig. 15 (although it is easier to see in a physical ball-and-stick model). Amorphous silicon is a technologically important material, as it can make a very inexpensive and moderately efficient solar-cell material [13]. Actually, despite the fact that theory indicates that a tetrahedrally-bonded network can be defect free even without a trace of long-range crystalline order, in practice pure amorphous silicon forms with a significant number of broken bonds (i.e., places in the network where the Si atoms are only three-fold coordinated. However, when hydrogenated, the material becomes a quite good semiconductor, and it is this that is used in photovoltaics. Evidently the H atoms can find the unsatisfied Si bonds and assure that all valence electrons are stably tied up in covalent bonds.

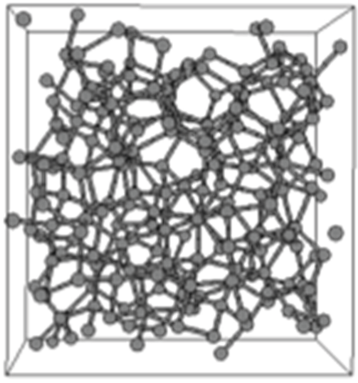


Fig. 15 *Theoretical construct of a random dense network with tetrahedral bonding.*

Of the enormous number of multi-component amorphous materials I will discuss only a few. The oxide of silicon SiO_2 that we saw earlier in crystalline (quartz) form in Fig. 8 occurs very readily in amorphous form – see Fig. 15 for a theoretical depiction of the sort of continuous covalently-bonded network that probably captures the essence of the atomic structure. One can

point to the large freedom offered by the paradigm of filling space with vertex-sharing tetrahedra, plus the floppiness of the O bond angle that we noted for quartz above, as reasons why this oxide so readily enters the amorphous state.

Finally, I note that many metal alloys can be amorphous. Apparently, the sphere packing problem with spheres of different sizes is not one that the atoms quickly "solve" upon solidification, so that disordered packings are not rare. On the other hand, for binary alloys crystallization is only avoided by giving the atoms very little time to find their best positions – this is achieved by freezing molten metal by shooting a stream of molten material on a cold rotating wheel. This is the method of "splat cooling". On the other hand, researchers have investigated metallic melts of, e.g., 7 different atomic species [14]. For these mixes the orderly sphere problem is apparently so difficult that they are found almost impossible to crystallize, so that even slow-cool castings of these strange steels are amorphous! Some unique advantages are seen for metal structures made from such very homogeneous, isotropic, and grain-boundary-free solids.

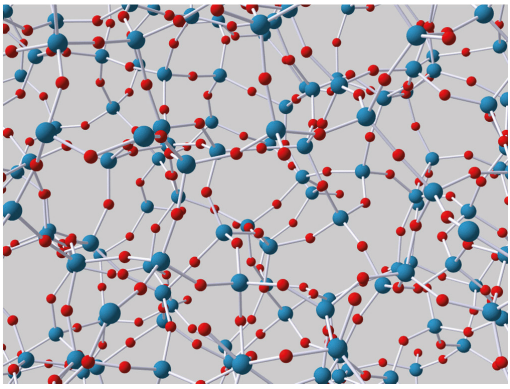


Fig. 16: *Theoretical construct of the amorphous network realized by* SiO_2.

6 Finally

I hope you have enjoyed this sampler from the wonderful, varied world of atomic structure in solids. Enjoy the rest of the school!

References

[1] To go a little bit off track immediately, I must add a proviso: paradoxically, ordinary crystallization is possible even when atoms have only repulsive interactions, if the system is under pressure; see B.J. Alder, T.E. Wainwright, "Phase transition for a hard sphere system." J. Chem. Phys. **27**:1208–1209 (1957).

[2] U. Mueller, “Inorganic Structural Chemistry” (2nd Edition, Wiley, 2007).

[3] “Basic Notions of Condensed Matter Physics”, P.W. Anderson, The Benjamin / Cummings Publishing Company, Inc., Advanced Book Program, Menlo Park, California, p. 12ff.

[4] V. Elser and C. L. Henley, “Crystal and Quasicrystal Structures in Al-Mn-Si Alloys”, Phys. Rev. Lett. **55**, 2883 (1985).

[5] D. Shechtman, I. Blech, D. Gratias, and J. W. Cahn, "Metallic Phase with Long-Range Orientational Order and No Translational Symmetry", Phys. Rev. Lett. **53**, 1951 (1984).

[6] R. Penrose, "The Role of Aesthetics in Pure and Applied Mathematical Research", Bull. Inst. Math. and its Appl. **10**, 266 (1971).

[7] While the word "atom" comes to us from the ancient Greeks, the history of the word "ion" is very different: it is a word made up to sound like it comes from ancient Greek, produced by 19th century Cambridge don William Whewell, who was asked to do it by Michael Faraday. Faraday understood that a whole suite of new words was necessary to express the revolutionary insights that came from his discoveries in electrochemistry. Together they coined the whole suite of words: ionic, ionize, anion, cation, anode, cathode, etc. Remarkably, Faraday did not even accept that atomic theory of matter! See L. P. Williams, “Michael Faraday – A Biography” (Basic Books, 1965), p. 262ff.

[8] K. Z. Rushchanskii, N. A. Spaldin, and M. Ležaić, "First-principles prediction of oxygen octahedral rotations in perovskite-structure EuTiO3," Phys. Rev. B **85**, 104109 (2012).

[9] H. F. Zhan, Y. Y. Zhang, J. M. Bell and Y. T. Gu, "Thermal conductivity of Si nanowires with faulted stacking layers," Journal of Physics D **47**, 015303 (2014).

[10] D. P. DiVincenzo, O. L. Alerhand, M. Schlüter, and J. W. Wilkins, “Electronic and Structural Properties of a Twin Boundary in Si,” Phys. Rev. Lett. **56**, 1925 (1986).

[11] W. T. Read, Jr., “Dislocations in Crystals” (McGraw-Hill, New York, 1953).

[12] S. Bysakha, K. Mitsuishi, M. Song, and K. Furuya, "Formation of amorphous xenon nanoclusters and microstructure evolution in pulsed laser deposited Ti(62.5)Si(37.5) thin films during Xe ion irradiation", J. Mater. Res. **26**, 62-69 (2011).

[13] “Amorphous Semiconductors”, ed. M. H. Brodsky (Springer Topics in Applied Physics Vol. 36, 1985).

[14] V. Ponnambalam, S. J. Poon, G. J. Shifle, "Fe–Mn–Cr–Mo–(Y, Ln)–C–B (Ln= Lanthanides) bulk metallic glasses as formable amorphous steel alloys," J. Mater. Res. **19**, 3046-3052 (2004).

A 2 Electronic Structure of Matter

Stefan Blügel and Gustav Bihlmayer
Peter Grünberg Institut and
Institute for Advanced Simulation
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Electrons in a periodic lattice **3**
2.1 Translation symmetry 3
2.2 Nearly free electrons 7
2.3 Bandstructures of selected systems 9

3 Interacting electrons **11**
3.1 The Hartree and the Hartree-Fock approximation 12
3.2 Density functional theory 15
3.3 Extensions to DFT: the LDA+U method 20
3.4 Quasiparticles and the GW approximation 23
3.5 Short summary: Calculating electronic structure of transition-metal oxides . . . 26

4 Relativistic effects **27**
4.1 Spin-orbit coupling 28
4.2 The Rashba- and the Dresselhaus effect 30
4.3 Topological insulators 33

1 Introduction

"Band theory" of solids goes all the way back to the theses of Felix Bloch [1], the first student of Werner Heisenberg in Leipzig, and Hans Bethe [2], student of Arnold Sommerfeld in Munich, who investigated the nature of independent non-interacting electrons in a periodic potential. These investigations culminated in the formulation of the concept of bands in crystals based upon a theorem what has come to be known as the "Bloch theorem", i.e. that the wavefunction in a perfect crystal is an eigenstate of the "crystal momentum". This was the starting point of a development, best described as the investigation of the electronic structure of matter, for which currently no end is in sight. Band theory plus Pauli exclusion principles allowed only states of one spin per one electron per unit cell of the crystal. The importance of filled electron states and empty "hole" states was recognized. The foundation was laid for the classification of materials into metals, semiconductors and insulators upon the number of electrons:

- Insulators have filled bands with a large energy gap of forbidden energies separating the ground state from all excited states of the electrons.
- Semicondutors have only a small gap, so that thermal energies are suffcient to excite the electrons to a degree that allows important conduction phenomena.
- Metals have partially filled bands with no excitation gaps, so that electrons can conduct electricity a zero temperature.

More than eighty years of research on the electronic structure of matter brought much progress along many different directions: Electron-electron interaction, disorder, topology, quantitative theories, to name a few. Electronic conduction in a solid is usually mediated by more or less extended states that allow charge carriers to travel in the lattice. If a material, for example a transition metal oxide like TiO_2 or a main group chalcogenide like $Ge_2Sb_2Te_5$, changes its electronic structure e.g. upon crystal transformation, this will be reflected by the charge carrying states of the material. Lattice imperfections, such as defects, or structural disorder modify the transport properties. Even though the conduction mechanisms can in the end be quite different, e.g. band conduction, topologically protected edge currents, polaronic transport, or trap-assisted tunneling, and there is a variety of insulating states like band-, Mott-, and Anderson-insulators, all described by specialized theories, the underlying picture of electronic states in a periodic crystal, the description by momentum, spin, and band index are fundamental concepts that are used in many of the refined theories.
In all these years "band-theory" became an indispensable tool for the description and characterization for many states of condensed matter, not only for (band) insulators, semiconductors or metals. Also in so-called strongly correlated electron systems, where a subset of states shows correlation effects that are typically not covered by band-theory, a proper description of the more delocalized bands is indispensable since the unique material properties are determined by the coupling of the localized, correlated states to these bands.
While these theories started as more conceptual tools for the understanding of the solid state, in the past 20 years advances both in theory and (computer) simulation techniques made it possible to predict electronic (and magnetic) properties on a quantum mechanical basis. Most notably, density functional theory developed into a reliable tool for material scientists, which aim at designing materials with selected properties. It allows, in some limits, to describe the influence of structural changes on the electronic system and to estimate the consequences for the conductive properties.

This lecture has an introductory character. The lecture tries to draw a line from the elementary concepts of band-theory of independent non-interacting electrons in a periodic potential to the modern theoretical frame work of density functional theory applied to complex solids, that provides a quantitative theory often of predictive power. To achieve this goal, necessarily a selection of topics is required. Some subjects are covered in some depth, others are included to put things into a wider perspective and provide an overview. Concerning the selection of materials, we have an emphasis on oxides and chalcogenides, relevant for resistive memories and phase change materials.
In a first section we will look at the basic properties of electrons in an infinite periodic lattice. We will ignore their mutual interactions, but incorporate the proper symmetry that defines the quantum numbers (constants of motion) of the system in a non-relativistic context. The interaction between electrons is then the topic of the second section, where methods will be discussed to treat Coulomb- and exchange interactions e.g. Hartree-Fock and density functional theory (DFT) or the GW approximation. As a simple combination of model-theory to describe strong correlation effects and DFT, we also introduce the DFT+U method here. Finally, we shortly touch the topic of topological insulators as it plays an interesting role in some phase change materials. In this context, we also introduce spin-orbit coupling that is important for a correct description of the more heavy elements in these materials.
Of course there are many materials – often with promising functional properties – that need a theoretical description beyond the one given in this lecture. Some of the subsequent lectures will give more specific insights, in particular about the description of transport and correlation aspects. Here, we deal mainly with the electronic ground state of matter from an "weakly correlated" point of view and we focus on periodic solids. Also discussion of solids with disorder or magnetic phenomena will be left for further lectures except for some illustrative examples.

2 Electrons in a periodic lattice

The electronic properties of a periodic solid are to a large part determined by the symmetry of the lattice that is formed by the atomic nuclei. They create a periodic potential in which the electrons (in particular the most loosely bound valence and conduction electrons) are moving. The constants of motion of such a system are determined by the symmetry, here in particular the translation symmetry in the crystal. Therefore, we start with a discussion of the symmetry properties of a crystal that leads us to Bloch's theorem and illustrate these considerations for the case of a nearly free electron gas. To be specific, we present a couple of prototypical band structures to see how – even in the presence of electron-electron interactions – many properties can be inferred from the crystal symmetry. For a more extended treatment of these topics, the reader is referred to standard textbooks on solid state physics, e.g. Ref. [3].

2.1 Translation symmetry

The structure of an infinite periodic crystal can be considered as a space filling repetition of non-overlapping units cells in three dimensions. The origin, $\mathbf{R}$, of an unit cell can be written as

$$\mathbf{R}_\mathbf{n} = \underline{A}\,\mathbf{n} \quad ; \quad \text{where} \quad \underline{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad \text{and} \quad \mathbf{n} = \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix} \tag{1}$$

where the n_i are positive or negative integer numbers labeling the unit cell and $\underline{A}$ is the Bravais matrix of the crystal. In each unit cell $\mathbf{n}$, a finite number of atoms (denoted by α) are located at positions

$$\mathbf{r}_{\mathbf{n},\alpha} = \mathbf{R}_\mathbf{n} + \boldsymbol{\tau}_\alpha \tag{2}$$

and the vectors $\boldsymbol{\tau}_\alpha$ are called the basis of the lattice. A crystal, that can be described by equations (1) and (2), is invariant under an infinite set of symmetry operations, which can be classified as translations, $\mathcal{T}$, and (proper and improper) rotations, $\mathcal{R}$, and combinations of these two. Here, we will focus on the translations, which act on some function in real space, $f(\mathbf{r})$:

$$\mathcal{T}_{\mathbf{R}_\mathbf{n}} f(\mathbf{r}) = f(\mathbf{r} + \mathbf{R}_\mathbf{n})\,. \tag{3}$$

The Hamiltonian of the electrons in a periodic solid consists of three parts: the kinetic energy of the electrons, T, their mutual Coulomb repulsion, $V_{\mathrm{e-e}}$ and the potential created by the nuclei, V_{ext}. The first two parts of the Hamiltonian are invariant with respect to any translation, but the latter term will only be unchanged if the translation vector is a lattice vector, $\mathbf{R}_\mathbf{n}$:

$$\mathcal{T}_{\mathbf{R}_\mathbf{n}} V_{\mathrm{ext}}(\mathbf{r}) = V_{\mathrm{ext}}(\mathbf{r} + \mathbf{R}_\mathbf{n}) = V_{\mathrm{ext}}(\mathbf{r})\,. \tag{4}$$

Therefore, the total Hamiltonian, $\mathcal{H}$, commutes with the translation operator $\mathcal{T}_{\mathbf{R}_\mathbf{n}}$ and both operators will have common eigenfunctions.
To find the eigenvalues γ of the translation operator, we consider the successive action of two translation operators on a function:

$$\begin{aligned} \mathcal{T}_{\mathbf{R}_{\mathbf{n}'}} \mathcal{T}_{\mathbf{R}_\mathbf{n}} f(\mathbf{r}) &= \mathcal{T}_{\mathbf{R}_{\mathbf{n}'}} \gamma(\mathbf{R}_\mathbf{n}) f(\mathbf{r}) = \gamma(\mathbf{R}_{\mathbf{n}'})\gamma(\mathbf{R}_\mathbf{n}) f(\mathbf{r}) \\ \mathcal{T}_{\mathbf{R}_{\mathbf{n}'}} \mathcal{T}_{\mathbf{R}_\mathbf{n}} f(\mathbf{r}) &= \mathcal{T}_{\mathbf{R}_{\mathbf{n}'}+\mathbf{R}_\mathbf{n}} f(\mathbf{r}) = \gamma(\mathbf{R}_{\mathbf{n}'} + \mathbf{R}_\mathbf{n}) f(\mathbf{r})\,. \end{aligned} \tag{5}$$

The fact that $\gamma(\mathbf{R}_{\mathbf{n}'})\gamma(\mathbf{R}_\mathbf{n}) = \gamma(\mathbf{R}_{\mathbf{n}'}+\mathbf{R}_\mathbf{n})$ suggests, that the vectors $\mathbf{R}$ appear in an exponential form in γ, i.e.

$$\gamma(\mathbf{R}_\mathbf{n}) = \mathrm{e}^{\mathbf{R}_\mathbf{n}\cdot\mathbf{P}}\,. \tag{6}$$

If we consider, that f is a normalized function, $\phi(\mathbf{r})$, that should not grow or vanish exponentially in an infinitely extended solid by applying a translation, we can assume that $\mathbf{P}$ is an imaginary quantity and write it as $i\mathbf{k}$. We can use this vector $\mathbf{k}$ to label the functions ϕ according to

$$\mathcal{T}_{\mathbf{R}_\mathbf{n}} \phi_\mathbf{k}(\mathbf{r}) = \mathrm{e}^{i\mathbf{R}_\mathbf{n}\cdot\mathbf{k}} \phi_\mathbf{k}(\mathbf{r})\,. \tag{7}$$

The matrix elements of the Hamilton operator with two such functions should be invariant to a lattice translation, i.e.

$$\begin{aligned} \langle \phi_{\mathbf{k}'}(\mathbf{r}) | \mathcal{H} | \phi_\mathbf{k}(\mathbf{r}) \rangle &= \langle \mathcal{T}_{\mathbf{R}_\mathbf{n}} \phi_{\mathbf{k}'}(\mathbf{r}) | \mathcal{H} | \mathcal{T}_{\mathbf{R}_\mathbf{n}} \phi_\mathbf{k}(\mathbf{r}) \rangle = \left\langle \mathrm{e}^{i\mathbf{R}_\mathbf{n}\cdot\mathbf{k}'} \phi_{\mathbf{k}'}(\mathbf{r}) | \mathcal{H} | \mathrm{e}^{i\mathbf{R}_\mathbf{n}\cdot\mathbf{k}} \phi_\mathbf{k}(\mathbf{r}) \right\rangle \\ &= \mathrm{e}^{i\mathbf{R}_\mathbf{n}\cdot(\mathbf{k}-\mathbf{k}')} \langle \phi_{\mathbf{k}'}(\mathbf{r}) | \mathcal{H} | \phi_\mathbf{k}(\mathbf{r}) \rangle \end{aligned} \tag{8}$$

which means, that either the exponential factor is unity or the matrix element must vanish. From the translation symmetry properties of the lattice we can thus conclude, that we only get non-vanishing matrix elements, if $\mathbf{R}_\mathbf{n} \cdot (\mathbf{k} - \mathbf{k}') = 2\pi N$, if N is some integer number. The latter condition can be brought into a slightly different form, if we write

$$\mathbf{R}_\mathbf{n} = \underline{A}\,\mathbf{n} \quad ; \quad \mathbf{k} - \mathbf{k}' = \underline{B}\,\mathbf{m} \quad \text{and} \quad \underline{A}\,\underline{B} = 2\pi\underline{1} \tag{9}$$

where $\underline{1}$ is the 3×3 unit matrix and $\underline{B}$ defines a lattice, where the lattice vectors are given by $\mathbf{K_m} = \underline{B}\,\mathbf{m}$ for integer vectors $\mathbf{m}$. This lattice is called the reciprocal lattice and $\mathbf{K_m}$ is called a reciprocal lattice vector. To get non-vanishing matrix elements, $\mathbf{k} - \mathbf{k}'$ must be a reciprocal lattice vector.
This result will help us in two ways: firstly, this result holds for all kinds of operators, that commute with the translation operator, i.e. $\langle\phi_{\mathbf{k}'}|\mathcal{O}|\phi_{\mathbf{k}}\rangle$ is only non-vanishing, if $\mathbf{k} - \mathbf{k}' = \mathbf{K_m}$. E.g. if $\mathcal{O}$ describes some excitation of the crystal and the ϕ's are wavefunctions of the ground- and excited state, we can derive selection rules from this symmetry. Secondly, if we consider that $\phi_{\mathbf{k}}$ is a trial function for the solution of the Dirac or the Schrödinger equation, we can immediately block-diagonalize the Hamiltonian in blocks of wavefunctions, where the difference of two wavevectors $\mathbf{k}$ and $\mathbf{k}'$ is a reciprocal lattice vector. This allows us to restrict the values of $\mathbf{k}$ to the smallest ones in each block and to use this vectors as quantum numbers that label the wavefunctions in the solid. The volume filled by these $\mathbf{k}$-vectors is called Brillouin zone and it is the equivalent of the Wigner-Seitz cell in real space, but now in reciprocal space.
The reciprocal lattice is particularly useful to describe lattice periodic functions:

$$u(\mathbf{r}) = \sum_{\mathbf{K_m}} \mathrm{e}^{i\mathbf{K_m}\cdot\mathbf{r}} u(\mathbf{K_m}) \quad \text{since} \quad u(\mathbf{r}+\mathbf{R_n}) = \sum_{\mathbf{K_m}} \mathrm{e}^{i\mathbf{K_m}\cdot(\mathbf{r}+\mathbf{R_n})} u(\mathbf{K_m}) = u(\mathbf{r})\,. \tag{10}$$

We can now write the eigenfunctions of the translation operator according to equation (7) as

$$\mathcal{T}_{\mathbf{R_n}}\phi_{\mathbf{k}}(\mathbf{r}) = \mathcal{T}_{\mathbf{R_n}}(\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r})) = \mathrm{e}^{i\mathbf{k}\cdot\mathbf{R_n}}\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}) = \mathrm{e}^{i\mathbf{k}\cdot\mathbf{R_n}}\phi_{\mathbf{k}}(\mathbf{r})\,. \tag{11}$$

Functions of the form $\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r})$ are called Bloch functions. They are the eigenfunctions of the translation operator $\mathcal{T}_{\mathbf{R_n}}$ and play an important role in the electron theory of periodic solids. Since we know that $\mathcal{T}_{\mathbf{R_n}}$ and the Hamiltonian commute, also the eigenfunctions of $\mathcal{H}$ can be written in this form:

$$\mathcal{H}\phi_{\mathbf{k},\nu}(\mathbf{r}) = \varepsilon_{\mathbf{k},\nu}\phi_{\mathbf{k},\nu}(\mathbf{r}) \quad ; \quad \phi_{\mathbf{k},\nu}(\mathbf{r}) = \mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k},\nu}(\mathbf{r})\,. \tag{12}$$

This is called Bloch's theorem. We introduced an additional quantum number ν to distinguish the different solutions that belong to the same vector $\mathbf{k}$. They will correspond to different values of $u_{\mathbf{k},\nu}(\mathbf{K_m})$ in equation (10).
Let us illustrate these points with the simplest possible example, a non-interacting electron gas in an uniform potential, V_0. In this case the many-electron wavefunction is separable into a product of single-particle wavefunctions and thus it is sufficient to study the Hamiltonian for a single particle, that is of the form

$$\mathcal{H} = -\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2 + V_0\,. \tag{13}$$

Omitting the constant potential and using atomic units ($\hbar = 1$, $m_\mathrm{e} = 1$) we can write the Schrödinger equation with Eq. (10) and Eq. (12) as

$$\begin{aligned} -\frac{1}{2}\boldsymbol{\nabla}^2\left(\sum_{\mathbf{K_m}} \mathrm{e}^{i(\mathbf{k}+\mathbf{K_m})\cdot\mathbf{r}} u_{\mathbf{k},\nu}(\mathbf{K_m})\right) &= \\ \frac{1}{2}\sum_{\mathbf{K_m}} (\mathbf{k}+\mathbf{K_m})^2 \mathrm{e}^{i(\mathbf{k}+\mathbf{K_m})\cdot\mathbf{r}} u_{\mathbf{k},\nu}(\mathbf{K_m}) &= \varepsilon_{\mathbf{k},\nu}\sum_{\mathbf{K_m}} \mathrm{e}^{i(\mathbf{k}+\mathbf{K_m})\cdot\mathbf{r}} u_{\mathbf{k},\nu}(\mathbf{K_m})\,. \end{aligned} \tag{14}$$

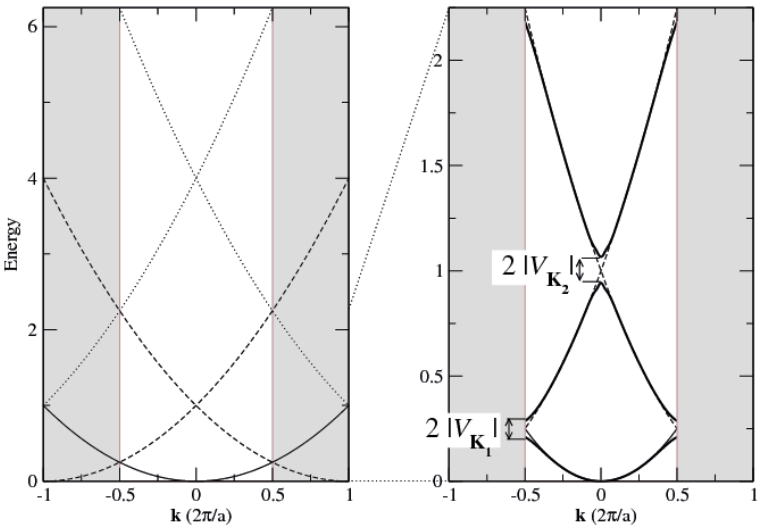


Fig. 1: *Band structure of a free electron gas (left) with three parabolic bands, originating from reciprocal lattice points outside the Brillouin zone (BZ, white). In the case of a nearly-free electron gas, i.e. in the presence of a periodic potential, the degeneracy of the bands at the BZ boundaries and at the origin will be lifted (thick lines, right panel).*

To fulfill Eq. (14) for each $\mathbf{K_m}$ we get a solution

$$\varepsilon_{\mathbf{k},\nu} = \frac{1}{2}(\mathbf{k}+\mathbf{K_m})^2\,, \tag{15}$$

i.e. the eigenvalues can be described as parabolas in $\mathbf{k}$-space originating at reciprocal lattice points. This is illustrated in figure 1 in one dimension: if we restrict our description to the first Brillouin zone (BZ), we observe that for each $\mathbf{k}$-vector we obtain an infinite, but discrete set of eigenvalues. At each $\mathbf{k}$-point, we can label these eigenvalues with the band index ν increasing with energy. A set of eigenvalues with the same index ν is called a band. At the boundaries of the BZ we observe band crossings, which will – in general – disappear for more realistic potentials (see next subsection).

Before we study the effect of a non-constant potential, it is instructive to study the problem of the free electron gas with a different choice for the wavefunctions. Although planewaves are a perfect solution to the free electron problem, in many real situations the eigenfunctions can be regarded to be derived from atomic wavefunctions, $s, p, d,$ or f-like. Of course in a crystal they form linear combinations that have to fulfill Bloch's theorem, i.e. if we start from orthonormalized atomic functions $\chi(\mathbf{r})$ centered on lattice sites $\mathbf{R_n}$, we choose a form

$$\phi_{\mathbf{k}}(\mathbf{r}) = \mathrm{e}^{i\mathbf{kr}} \sum_{\mathbf{n}} \chi(\mathbf{r}-\mathbf{R_n})\,. \tag{16}$$

E.g. if χ is a spherical s-like function, it is modulated by the $\mathbf{k}$-dependent "Bloch factor" with a period of $2\pi/k$ throughout the crystal. In the spirit of the tight-binding approximation, it is

convenient to write

$$\phi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{N}} \sum_{\mathbf{n}} \mathrm{e}^{i\mathbf{k}\mathbf{R}_{\mathbf{n}}} \chi(\mathbf{r} - \mathbf{R}_{\mathbf{n}}) \,, \tag{17}$$

where N is the number of lattice sites in the sum. This allows us to estimate the energy as the expectation value of the Hamiltonian:

$$\varepsilon(\mathbf{k}) = \frac{1}{N} \sum_{\mathbf{n},\mathbf{n}'} \left\langle \mathrm{e}^{i\mathbf{k}\mathbf{R}_{\mathbf{n}}} \chi(\mathbf{r} - \mathbf{R}_{\mathbf{n}}) | \mathcal{H} | \mathrm{e}^{i\mathbf{k}\mathbf{R}_{\mathbf{n}'}} \chi(\mathbf{r} - \mathbf{R}_{\mathbf{n}'}) \right\rangle = \sum_{\mathbf{n}} \mathrm{e}^{i\mathbf{k}\mathbf{R}_{\mathbf{n}}} \left\langle \chi(\mathbf{r}) | \mathcal{H} | \chi(\mathbf{r} - \mathbf{R}_{\mathbf{n}}) \right\rangle \,. \tag{18}$$

Assuming a linear chain of atoms with a lattice constant a, where only the nearest neighbor atoms have significant overlap, this reduces to

$$\varepsilon(\mathbf{k}) = \left\langle \chi(\mathbf{r}) | \mathcal{H} | \chi(\mathbf{r}) \right\rangle + \mathrm{e}^{i\mathbf{k}\mathbf{a}} \left\langle \chi(\mathbf{r}) | \mathcal{H} | \chi(\mathbf{r} - \mathbf{a}) \right\rangle + \mathrm{e}^{-i\mathbf{k}\mathbf{a}} \left\langle \chi(\mathbf{r}) | \mathcal{H} | \chi(\mathbf{r} + \mathbf{a}) \right\rangle \,. \tag{19}$$

Since the nearest-neighbor integrals are identical, we can write

$$\varepsilon(\mathbf{k}) = \alpha + 2\beta \cos(ka) \quad \text{where} \quad \alpha = \left\langle \chi(\mathbf{r}) | \mathcal{H} | \chi(\mathbf{r}) \right\rangle \quad \text{and} \quad \beta = \left\langle \chi(\mathbf{r}) | \mathcal{H} | \chi(\mathbf{r} \pm \mathbf{a}) \right\rangle \,. \tag{20}$$

For a s-type wavefunction, β is negative and ε is lowest at $k = 0$. The Bloch phase in front of each atomic orbital is positive and the wave function describes a bonding state. With increasing k the energy increases and reaches its maximum at $k = \pi/a$. At this k-point, the Bloch factor changes sign at every second atom and the wavefunction describes an antibonding state. For a p-type wavefunction, which is an "odd" functions, β is positive and their energy decreases from $k = 0$ towards the zone boundary, like indicated in the second band in figure 1. As we will see later, realistic bandstructures of simple metals indeed start at low energies with a parabolic, s-type band, followed by three inverted parabolas that correspond to p-type states. Many properties of the valence electrons of these metals can be found in the simple free-electron picture of this section.

2.2 Nearly free electrons

Of course, the electrons in a crystal feel a periodic potential that differs considerably from our constant model potential. The attractive potential of the nuclei is partially screened by energetically low-lying core electrons, i.e. the valence electrons "feel" a potential that is in essence smoothed by electrons that are bound closely by the nuclei. This is the external potential, V_{ext}, encountered in Eq. (4). Since it has lattice periodicity, it can also be expanded in reciprocal lattice vectors:

$$V_{\mathrm{ext}} = \sum_{\mathbf{K}_{\mathbf{m}}} \mathrm{e}^{i\mathbf{K}_{\mathbf{m}}\cdot\mathbf{r}} V(\mathbf{K}_{\mathbf{m}}) \,. \tag{21}$$

Adding this potential term to the Hamiltonian modifies Eq. (14) and we get

$$\sum_{\mathbf{K}_{\mathbf{m}}} \left(\varepsilon_{\mathbf{k},\nu} - \frac{1}{2}(\mathbf{k} + \mathbf{K}_{\mathbf{m}})^2 \right) u_{\mathbf{k},\nu}(\mathbf{K}_{\mathbf{m}}) \mathrm{e}^{i(\mathbf{k}+\mathbf{K}_{\mathbf{m}})\mathbf{r}} = \sum_{\mathbf{K}_{\mathbf{m}}} \sum_{\mathbf{K}'_{\mathbf{m}}} V(\mathbf{K}'_{\mathbf{m}}) u_{\mathbf{k},\nu}(\mathbf{K}_{\mathbf{m}}) \mathrm{e}^{i(\mathbf{k}+\mathbf{K}_{\mathbf{m}}+\mathbf{K}'_{\mathbf{m}})\mathbf{r}}. \tag{22}$$

Introducing $\mathbf{K}''_{\mathbf{m}} = \mathbf{K}'_{\mathbf{m}} + \mathbf{K}_{\mathbf{m}}$, we write the right part of Eq. (22) as

$$\sum_{\mathbf{K}''_{\mathbf{m}}} \sum_{\mathbf{K}'_{\mathbf{m}}} V(\mathbf{K}'_{\mathbf{m}}) u_{\mathbf{k},\nu}(\mathbf{K}''_{\mathbf{m}} - \mathbf{K}'_{\mathbf{m}}) \mathrm{e}^{i(\mathbf{k}+\mathbf{K}''_{\mathbf{m}})\mathbf{r}}. \tag{23}$$

Substituting back $\mathbf{K_m} \leftarrow \mathbf{K''_m}$ and comparing the coefficients with the left side of Eq. (22) we obtain

$$\left(\varepsilon_{\mathbf{k},\nu} - \frac{1}{2}(\mathbf{k}+\mathbf{K_m})^2\right) u_{\mathbf{k},\nu}(\mathbf{K_m}) = \sum_{\mathbf{K'_m}} V(\mathbf{K'_m}) u_{\mathbf{k},\nu}(\mathbf{K_m} - \mathbf{K'_m})\,. \tag{24}$$

For the case of a constant potential, we set $V(\mathbf{0}) = V_0$ and all other Fourier coefficients to zero. In this case, Eq. (24) reduces to

$$\left(\varepsilon_{\mathbf{k},\nu} - \frac{1}{2}(\mathbf{k}+\mathbf{K_m})^2\right) u_{\mathbf{k},\nu}(\mathbf{K_m}) = V_0 u_{\mathbf{k},\nu}(\mathbf{K_m})\,, \tag{25}$$

which corresponds, apart from an additional constant V_0, to Eq. (15). Eigenfunctions are again planewaves with wave vector $\mathbf{K_m}$, i.e. the Fourier coefficients for the expansion of the wavefunction, $u_{\mathbf{k},\nu}(\mathbf{K_m})$ for a certain state ν are unity for a specific $\mathbf{K_m}$ and zero otherwise. If we consider $u_{\mathbf{k}}$ as a matrix with dimensions ν and $\mathbf{K_m}$, we find that for the case of a constant potential, $\underline{u_{\mathbf{k}}}$ is the unit matrix $\underline{1}$ (we denote matrix quantities here and in the following with an underline).

If the potential is of general shape, electrons cannot be described by eigenstates of a single reciprocal lattice vector, instead the expansion coefficients $u_{\mathbf{k},\nu}(\mathbf{K_m})$ can be obtained from Eqs. (24). We can rewrite these equations using $\frac{1}{2}(\mathbf{k}+\mathbf{K_m})^2 = \varepsilon^0_{\mathbf{k},\mathbf{K_m}}$ in the form

$$\left(\varepsilon_{\mathbf{k},\nu} - \varepsilon^0_{\mathbf{k},\mathbf{K_m}}\right) u_{\mathbf{k},\nu}(\mathbf{K_m}) = \sum_{\mathbf{K'_m}} V(\mathbf{K'_m} - \mathbf{K_m}) u_{\mathbf{k},\nu}(\mathbf{K'_m})\,. \tag{26}$$

If we write V in matrix form and consider u and ε^0 as vectors, this equation can be rewritten in the form of a standard eigenvalue problem:

$$\left(\underline{V} + \underline{\varepsilon^0_{\mathbf{k}}}\right) \mathbf{u}_{\mathbf{k},\nu} = \varepsilon_{\mathbf{k},\nu} \mathbf{u}_{\mathbf{k},\nu}\,. \tag{27}$$

Let us finally analyze, how a weakly varying potential affects the bandcrossings at $k = \pi/a$ in figure 1. The lowest bandcrossing is formed by a parabola originating at $\mathbf{K_m} = \mathbf{0}$ and a parabola that has its minimum at $\mathbf{K_m} = \mathbf{K}_1$. In the vicinity of the crossing, we denote the (unperturbed) eigenvalues of these two states as $\varepsilon^0_+ = V_0 + \varepsilon^0_{\mathbf{0}}$ and $\varepsilon^0_- = V_0 + \varepsilon^0_{\mathbf{K}_1}$. Ignoring all other states, the potential matrix $\underline{V}$ will be a 2×2 matrix, which has diagonal elements V_0 and off-diagonal elements $V_{-\mathbf{K}_1}$ and $V_{\mathbf{K}_1}$. Eq. (27) has then the form

$$\begin{pmatrix} V_0 + \varepsilon^0_{\mathbf{0}} & V_{\mathbf{K}_1} \\ V_{-\mathbf{K}_1} & V_0 + \varepsilon^0_{\mathbf{K}_1} \end{pmatrix} \begin{pmatrix} u_{\mathbf{0}} \\ u_{\mathbf{K}_1} \end{pmatrix} = \varepsilon \begin{pmatrix} u_{\mathbf{0}} \\ u_{\mathbf{K}_1} \end{pmatrix}\,. \tag{28}$$

This problem can be solved by setting the determinant to zero:

$$\begin{vmatrix} \varepsilon^0_+ - \varepsilon & V_{\mathbf{K}_1} \\ V_{-\mathbf{K}_1} & \varepsilon^0_- - \varepsilon \end{vmatrix} = 0 \tag{29}$$

resulting in

$$\varepsilon_{1,2} = \frac{\varepsilon^0_+ + \varepsilon^0_-}{2} \pm \sqrt{\left(\frac{\varepsilon^0_+ - \varepsilon^0_-}{2}\right)^2 + |V_{\mathbf{K}_1}|^2}\,. \tag{30}$$

In case the two eigenvalues ε^0_+ and ε^0_- coincide, the non-constant potential will lift this degeneracy and lead to a splitting of $\pm|V_{\mathbf{K}_1}|$. If $\varepsilon^0_+ - \varepsilon^0_-$ is large compared to $|V_{\mathbf{K}_1}|$, e.g. $\mathbf{k}$ is far from

the zone boundary, the effect will be small. Likewise, the interaction with other bands, that are energetically far away, will be small and our assumption to consider just two bands near a crossing will be justified.
Of course, in realistic bandstructures more than two bands can be energetically close and not all potential Fourier coefficients will be small, so that more complicated bandstructure scenario will occur. In this case, it is necessary to solve Eq. (27) in full.
An other important issue is the occupation of bands. For each Bloch vector $\mathbf{k}$ the number of bands are bound from below, but the eigenvalue spectrum has no upper bound. Electrons are fermions and no state defined by the quantum numbers, Bloch vector, band index and spin, $(\mathbf{k}\nu\sigma)$, can be occupied twice. We will show below that we occupy all states from below till all valence electrons of all atoms of an unit cell occupy an available state once. This typically entails that the unit cell is charge neutral. It implies the introduction of an important quantity, the Fermi energy, E_{F}, the energy of the highest occupied states at temperature $T = 0$ K. The surface that is created by all states $\mathbf{k}, \nu$, which fulfill the condition $\varepsilon_{\mathbf{k},\nu} = E_{\mathrm{F}}$, is known as the Fermi surface. The shape of the Fermi surface is a characteristic fingerprint of each metal. This surface separates the occupied from the unoccupied states and plays therefore an important role in all transport properties of metals. In metals conductance can occur at zero temperature for infinitesimally small external fields. In semiconductors and insulators the Fermi energy is placed in the gap and any excitation requires a finite energy of at least the size of the band-gap. Thermal excitations at finite temperature T, which take place around the Fermi energy are described by the Fermi-Dirac distribution $f(\varepsilon, T) = 1/[\exp((\varepsilon - E_{\mathrm{F}})/(k_{\mathrm{B}}T)) + 1]$, where $k_{\mathrm{B}} = 8.6173324(78) \times 10^{-5}$ eV/K is the Boltzmann constant.

2.3 Bandstructures of selected systems

To visualize the effects of the crystal lattice, we shortly discuss here two prototypical examples of bandstructures. First, we consider the sodium crystal, which crystallizes in the body-centered cubic lattice. The valence electron is only weakly bonded and contributes to a very low electron density. It is well described by the nearly-free electron model. In contrast, the perovskite lattice is strongly ionic and the p-states of oxygen and the d-states of the transition-metal are rather localized and these states are rather derived from the atomic levels than from a free-electron model. Nevertheless, their dispersion follows for small k-vectors the predictions of our simple theory.
First we discuss the electronic structure of sodium, which is a simple metal from the first column of the periodic table. All the metals in this row crystallize in a bcc lattice and with a single valence electron per atom. Thus, band is half-full, i.e. the highest occupied states cut through the bands, no gap exists, which explains why sodium and all other alkali metals are metals. Thus, the Na bandstructure is prototypical for these elements, like K or Rb [4]. Due to its low electron density (the s-electrons are typically very delocalized in metals), it is already very close to an almost free electron gas in a periodic lattice.
The bandstructure in figure 2 has been obtained by density functional theory (DFT), that will be outlined in the next section (3.2). As can be seen from this figure, the bottom of the occupied band is almost parabolic, as expected for a free-electron like dispersion. We see, however, that gaps are opening at the boundaries of the Brillouin zone, e.g. at the N-point. Above the s-band three downward dispersing p-bands can be observed with very different dispersions. E.g. in ΓN direction, only one band reaches down to the s-band, while the other two bands remain above 8 eV.

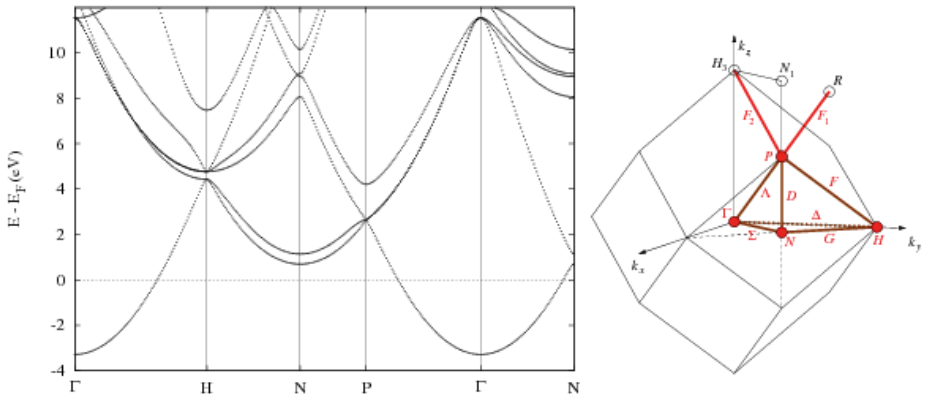


Fig. 2: *Band structure of sodium (left) and the reciprocal unit cell of the bcc lattice with high symmetry points and lines (right). The right image was taken from the Bilbao Crystallographic Server [5].*

We should notice, that even in this very simple case one has to be careful when comparing single-particle eigenvalues (figure 2) with experimental photoemission results: While the Fermi surface is in very good agreement, the bottom of the s-band is too low as compared to the experiment. Photoemission results show that 2.5 eV are required to excite an electron at the Γ-point to the Fermi level, while the DFT eigenvalue is at about -3.2 eV [6]. Methods to calculate excitation spectra will be shortly discussed in subsection 3.4.

As a second example, very different from the nearly-free electron case, we show in figure 3 the band structure of the perovskite $SrTiO_3$, again obtained by DFT. $SrTiO_3$ is an insulator. There is a gap in the eigenvalue spectrum for all $\mathbf{k}$ vectors in the Brillouin zone. The Fermi energy lies in the band gap. The occupied bands are called valence bands and the unoccupied ones are referred to as conduction bands. The maximum of the valence band and the minimum of the conduction band occur at a different $\mathbf{k}$ vectors. In this case the band-gap is called an indirect gap, in opposite to e.g. GaAs, which exhibits a direct gap, i.e. maximum of the valence band and the minimum of the conduction band coincide at the same $\mathbf{k}$ vector in the Brillouin zone, at the center of the Brillouin zone, i.e. $\mathbf{k} = \mathbf{0}$.

Although the situation is now more involved we can still interpret the result based on the model discussed above. For this it is convenient to introduce the concept of the density of states (DOS) as the number of states $g(\varepsilon)$ given in an energy interval $d\varepsilon$: Each band ν contributes to the DOS, $g(\varepsilon)$, as

$$g_\nu(\varepsilon)\, d\varepsilon = \frac{1}{4\pi^3} \int d\mathbf{k} \times \begin{cases} 1 \text{ for } \varepsilon \leq \varepsilon_\nu(\mathbf{k}) \leq \varepsilon + d\varepsilon \\ 0 \text{ otherwise} \end{cases} . \tag{31}$$

In addition, this quantity can be weighted according to the localization of the states in a specific area of the crystal, e.g. near some atom, to yield the so-called local DOS. This quantity gives a useful measure how many states of a certain character are available in some energy region. As can be seen from figure 3, the band structure of $SrTiO_3$ is dominated by occupied oxygen (p) states below the Fermi level and the unoccupied Ti (d) states above E_F. The nine bands between

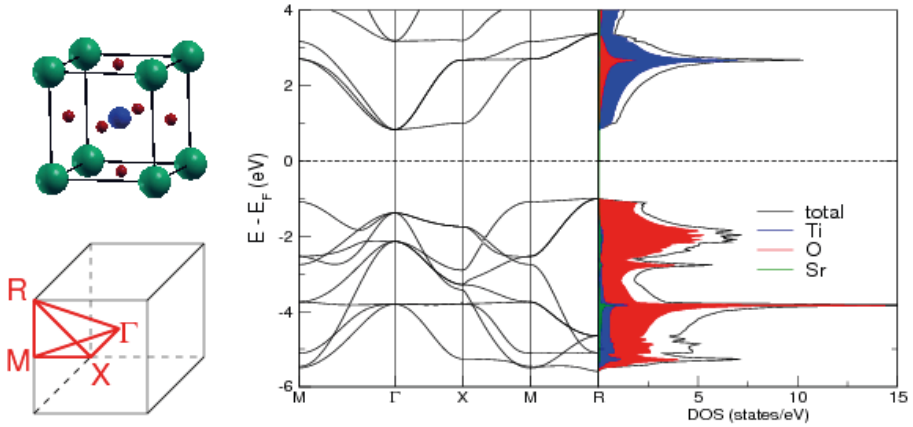


Fig. 3: *Crystal structure (upper left) and band structure (right) of $SrTiO_3$ along the path indicated at the lower left. The density of states (DOS) and the local DOS are shown on the right. Sr, Ti and O atoms are shown in green, blue and red, respectively.*

-6 and -1 eV result from the three O atoms, i.e. they are formed by p-like states as can be seen from their mostly downward-dispersing character at the Γ-point. Along a certain direction, e.g. k_x along $\Gamma - X$, only the p_x orbitals overlap enough to show strong dispersion, while p_y and p_z remain almost dispersionless. On the other hand, above 1 eV we recognize upward dispersing Ti d states that are split by the octahedral crystal field of the oxygen atoms into three t_{2g}-type levels (1 to 3 eV) and two e_g-type bands (above 3 eV). Also here, some bands remain flat if the orbitals are orthogonal to the direction in k-space (e.g. d_{yz} orbitals along k_x).
DFT predicts a gap between these O and Ti states of about 1.9 eV, significantly smaller than what is found in experiment (3.25 eV). We will turn to this point later, in the discussion of the GW approximation. Nevertheless, the character of the bands near E_F is quite reliably described by the DFT calculation.

3 Interacting electrons

In any realistic calculation, we cannot simply ignore the mutual Coulomb repulsion of the electrons, which is described by the term

$$V_{\mathrm{e-e}} = \frac{1}{2} \sum_{i \neq j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{32}$$

in the Hamiltonian (since we work in atomic units, $e^2 = 1$). This destroys the separability of the many-body wavefunction into single-particle wavefunctions and the straightforward quantum mechanical treatment of the electronic degrees of freedom is limited to a very small number of particles. This is mainly due to the appearance of the many-body wavefunction $\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$, which contains a tremendous amount of information and is difficult to handle for N larger than a few dozen or so.

One can try to construct Ψ from single-particle wavefunctions and combine them to many-body wavefunctions of different complexity: a simple product Ansatz, $\Psi = \phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)\ldots\phi_N(\mathbf{r}_N)$, leads to the so called Hartree approximation. This form of the wavefunction is, however, not compatible with the Pauli principle, i.e. interchanging two arguments of Ψ does not lead to $-\Psi$. In the Hartree-Fock (HF) method, the wavefunction has the form of a $N \times N$ matrix of single-particle wavefunctions $\phi_\mu(\mathbf{r}_\nu)$ with $1 \leq \mu, \nu \leq N$, which ensures that the Pauli principle is fulfilled. Therefore, the HF method leads to better results (e.g. binding energies) than the Hartree method. The energy contribution missing in the latter method as compared to the former one is called *exchange energy*. Although the HF method is numerically quite complicated, the obtained energies are still often quite far from the true ground state energies. What is missing is called *correlation energy* and the results can be improved by e.g. constructing the many-body wavefunction as a linear combination of many determinant functions. These so called configuration-interaction (CI) methods belong to the realm of quantum-chemical methods and can be systematically improved, but the numerical effort is huge. While the HF method scales nominally like N^4, calculational schemes that include correlation scale with N^5 (second order Møller-Plesset perturbation theory) or N^7 (Coupled Cluster theory). A good account of these quantum-chemical methods can be found in Ref. [7] and [8].
A completely different approach is taken by the density functional theory (DFT): although in most cases the true wavefunction is impossible to access, this poses no fundamental limitation since normally we are not interested in Ψ, but in a limited number of physical observables. Density functional theory therefore bypasses the troublesome many-body wavefunction and starts directly from the density of the particles in question (in our case electrons) allowing thereby the treatment of a large number of particles.
As we will see, DFT is quite suitable to describe the many aspects of the electronic structure of solids. Also structural properties, like lattice parameters, actually many quantities that can be obtained from total energies are very well accessible in DFT, since the total energy is a quantity that has a definite meaning in this theory. In metals also other electronic properties are reproduced well, mainly due to the fact that they are dominated by states near the Fermi level. Evidence, that these electrons can be qualitatively described in an independent particle description (similar to the "particles" in DFT) comes from Landau's Fermi liquid theory [3, 9]. To understand the approximations and models that enter DFT and its extensions, it is illustrative first to remember the maybe most basic theories that describe electron-electron interaction, the Hartree and the Hartree-Fock approximation to the solution of the many-body Schrödinger equation.

3.1 The Hartree and the Hartree-Fock approximation

Let us start with the many-body Schrödinger equation for the electrons

$$\sum_i \left(h_i + \frac{1}{2} \sum_{i \neq j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \right) \Psi = \varepsilon \Psi \qquad \text{with} \qquad h_i = -\frac{1}{2}\boldsymbol{\nabla}_i^2 + V_{\text{ext}}(\mathbf{r}_i)\,, \tag{33}$$

where $V_{\text{ext}}(\mathbf{r})$ includes the potential arising from the interaction with the nuclei and other possible external potentials. Assume that we found solutions to the single-particle Hamiltonian h_i and denote them $\phi_i(\mathbf{r}_i)$. Then, we can try to construct Ψ from these single-particle wavefunctions and combine them to a many-body wavefunctions by a simple product Ansatz, $\Psi = \phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)\ldots\phi_N(\mathbf{r}_N)$.

Ignoring the interaction part in Eq. (33) for the moment, we study a system of independent electrons. Then, if we multiply from the left with all single-particle wavefunctions except one (e.g. ϕ_i) and integrate over all $\mathbf{r}$'s except $\mathbf{r}_i$, we arrive at a set of equations

$$h_i\phi_i(\mathbf{r}_i) = (\varepsilon - \sum_j \epsilon_j)\phi_i(\mathbf{r}_i) \qquad \text{where} \qquad h_j\phi_j(\mathbf{r}_j) = \epsilon_j\phi_j(\mathbf{r}_j)\,. \tag{34}$$

In this case, the eigenvalue of the many-body wavefunction, ε, is obviously the sum of all single-particle eigenvalues, ϵ_i. Given the fact that electrons are fermions and cannot occupy a state more than once, this means that the ground state of our system will be the one that has the lowest N single-particle states occupied.
Now, if we reintroduce the electron-electron interaction and go though the same steps, we get a coupled set of equations

$$\left(h_i + V_i(\mathbf{r}_i)\right)\phi_i(\mathbf{r}_i) = (\varepsilon - \sum_j \epsilon_j)\phi_i(\mathbf{r}_i) \qquad \text{with} \qquad V_i(\mathbf{r}_i) = \sum_{j\neq i}\left\langle \phi_j(\mathbf{r}_j)\middle|\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}\middle|\phi_j(\mathbf{r}_j)\right\rangle , \tag{35}$$

where $V_i(\mathbf{r}_i)$ is the potential created by all electrons except the one described by ϕ_i. Solving these equations is already a complicated task, but a considerable simplification can be achieved if we assume that in an infinite solid there are so many electrons in the system, that we can assume that every electrons "sees" the same potential arising from all the states

$$V_{\mathrm{H}}(\mathbf{r}) = \sum_j \left\langle \phi_j(\mathbf{r}_j)\middle|\frac{1}{|\mathbf{r} - \mathbf{r}_j|}\middle|\phi_j(\mathbf{r}_j)\right\rangle . \tag{36}$$

This leaves a single equation for all states

$$\left(h + V_{\mathrm{H}}(\mathbf{r})\right)\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}) \tag{37}$$

which has to be solved self-consistently. This means, since V_{H} – the Hartree potential – depends on the states ϕ_i, first a guess for this potential has to be made (e.g. for states calculated in the independent electron approximation), and then Eq. (37) can be solved initially. With the solutions in the next iteration a new, better guess for V_{H} can be obtained and this process can be repeated until the potential does not change any more from one iteration to another.
What we can learn from this so-called Hartree method are two things: in some approximation we can retain the notion of single-particle states and occupy them by an Aufbau-principle to construct a many-body wavefunction. In a self-consistent scheme, equations for these single-particle states can be solved to obtain a solution iteratively. The other lesson to learn is that already a simple product Ansatz is very difficult to handle unless we make approximations, like substituting the state-dependent potential V_i by the Hartree potential, V_{H}, thereby introducing some self-interaction of the single-particle states.
However, this is not the most severe shortcoming of the Hartree method: as mentioned above, the biggest approximation we introduced in the beginning by choosing a simple product Ansatz. This construction of the many-body wavefunction of N non-interacting electrons still suffers from a serious problem in the treatment of the fermionic nature of the electrons. The simple product of single-particle states does not fulfill a basic requirement for fermions, which states that the many-body wavefunction has to be anti-symmetric under the exchange of two particles

$$\Psi(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_N) = -\Psi(\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N)\,, \tag{38}$$

where we introduced $\mathbf{x} = (\mathbf{r}, \sigma)$ to denote the combination of the spatial and spin degrees of freedom. For the moment it is sufficient to consider the spin, σ, simply as a label that can assume two values.
However, it was realized early by Slater [10], that an anti-symmetric linear combination of product wavefunctions can be constructed, which has the desired property. This construction is known as a 'Slater determinant' as it can be expressed in terms of a determinant of a matrix containing the single-particle states

$$\begin{aligned}\Psi_{\text{Slater}}(\mathbf{x}_1 \dots \mathbf{x}_N) &= \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\mathbf{x}_1) & \dots & \phi_1(\mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \phi_N(\mathbf{x}_1) & \dots & \phi_N(\mathbf{x}_N) \end{vmatrix} \\ &= \frac{1}{\sqrt{N!}} \sum_P (-1)^P P\left(\phi_1(\mathbf{x}_1) \dots \phi_N(\mathbf{x}_N)\right). \end{aligned} \tag{39}$$

In this notation the sum is performed over all permutations P acting on the indices i of the ϕ_i. The factor $(-1)^P$ ensures the required anti-symmetry.
The Slater-determinants as given in Eq. (39) form an anti-symmetric solution of the non-interacting Schrödinger equation. It can be shown that using all possible combinations of single-particle wavefunctions these determinants form a basis of the space of the N-body wavefunctions so that the interacting many-body wavefunction can be expressed as a linear combination of Slater determinants. Such an expansion forms the basis of complicated and expensive computational methods like so called configuration interaction calculations which yield high accuracy but can be performed only for a small number of electrons N.
Using these determinant functions to find a solution for the Hamiltonian as shown in Eq. (33) is the essence of the Hartree-Fock method. The derivation of the equations is somewhat lengthy but rather straightforward. The strategy is to vary the functions ϕ to make the expectation value of the many-body Hamiltonian $\langle\Psi|\mathcal{H}|\Psi\rangle$ an extremum under the constraint that the functions ϕ are normalized to unity. This constraint is introduced by an Lagrange multiplier $\varepsilon_{i,\sigma}$ for each $\phi_{i,\sigma}$. This leads then to the so called Hartree-Fock equation

$$\left(-\frac{1}{2}\boldsymbol{\nabla}^2 + V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r})\right)\phi_{i,\sigma}(\mathbf{r}) + \sum_{j,\sigma'} \int \frac{\phi^*_{j,\sigma'}(\mathbf{r}')\phi_{i,\sigma}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' \phi_{j,\sigma'}(\mathbf{r}) = \varepsilon_{i,\sigma}\phi_{i,\sigma}(\mathbf{r}). \tag{40}$$

It is the last term on the left side of Eq. (40) that introduces the physics missing in the Hartree method. It can be rewritten to give the equation a more familiar form:

$$\left(-\frac{1}{2}\boldsymbol{\nabla}^2 + V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r}) + V_{\text{ex}}(\mathbf{r}; i\sigma)\right)\phi_{i,\sigma}(\mathbf{r}) = \varepsilon_{i,\sigma}\phi_{i,\sigma}(\mathbf{r}). \tag{41}$$

In addition to the non-interacting single-particle Hamiltonian two additional single-particle potential terms appear, which describe the Coulomb interaction. The first one we know already as the Hartree potential due to the interaction of the electron charge interacting with all electron charges including itself via the classical Coulomb interaction. The second term, the so called exchange potential, is a combined effect of Coulomb interaction and the antisymmetry condition of Fermions and can, therefore, not be interpreted classically. This term can be written

$$V_{\text{ex}}(\mathbf{r}; i\sigma) = -\frac{1}{\phi^*_{i,\sigma}(\mathbf{r})\phi_{i,\sigma}(\mathbf{r})} \sum_{j,\sigma'} \left\langle \phi_{i,\sigma}(\mathbf{r})\phi_{j,\sigma'}(\mathbf{r}')| \frac{1}{|\mathbf{r}-\mathbf{r}'|} |\phi_{i,\sigma}(\mathbf{r}')\phi_{j,\sigma'}(\mathbf{r}) \right\rangle', \tag{42}$$

where the integration ($\langle\rangle'$) is assumed over $\mathbf{r}'$. If the summation in Eq. (42) would be restricted to the term $j = i$ only, we would recover Eq. (35) indicating that the exchange potential in the Hartree-Fock method contains a self-interaction correction, making the Harfree-Fock theory self-interaction free. Again we arrive at a state dependent potential, that can be thought to originate from a (nonlocal) charge density

$$n_{\text{ex}}^{i\sigma}(\mathbf{r},\mathbf{r}') = \sum_{j,\sigma'} \frac{\phi_{i,\sigma}^*(\mathbf{r})\phi_{j,\sigma'}^*(\mathbf{r}')\phi_{i,\sigma}(\mathbf{r}')\phi_{j,\sigma'}(\mathbf{r})}{\phi_{i,\sigma}^*(\mathbf{r})\phi_{i,\sigma}(\mathbf{r})}\,. \tag{43}$$

n_{ex} has the property that it integrates to unity, so it corresponds to the charge of a single electron. Furthermore, for a spin σ and the limit $\mathbf{r} = \mathbf{r}'$ it reduces to the state-independent value $\sum_j \phi_{j,\sigma}^*(\mathbf{r})\phi_{j,\sigma}(\mathbf{r})$. This charge density is also called the exchange hole, describing the influence of a state i,σ when moving through the ensemble of all states in the system. We will encounter the exchange hole once more in the context of density functional theory where it appears in a state-independent form, actually very similar to Slater's idea [11] of a state-averaged version of Eq. (43) that inspired also the conception of the first exchange-correlation potentials for DFT.
To describe the electronic properties of metals, semiconductors or even insulators the Hartree-Fock theory, as presented here, is usually not the best choice. Although structural parameters might be reasonable, the missing electron correlation leads e.g. to strongly overestimated band gaps - for $SrTiO_3$ more than 12 eV are found [12]. Nevertheless, the fact that exchange is described exactly in this theory led to new developments that combine orbital-dependent terms [in the spirit of Eq. (42)] with correlation as described in conventional DFT functionals [13]. So-called hybrid functionals gained considerable popularity in particular for the calculation of perovskites [14]. To understand how these methods work, we first have to discuss the basic principles of DFT.

3.2 Density functional theory

While many researchers were working on more tractable versions of the Hartree-Fock method, in the middle of the sixties Hohenberg and Kohn [15] worked out two central theorems that form the basis of a conceptually different approach, the density functional theory: Consider a system of N particles (e.g. electrons) moving in an external potential $V(\mathbf{r})$ (caused by e.g. nuclei). In a non-degenerate ground state (i) the many-body wavefunction Ψ and $V(\mathbf{r})$ are uniquely determined by the particle density distribution $n(\mathbf{r})$ and (ii) there exists an energy functional of this density, $E[n(\mathbf{r})]$, which is stationary with respect to variations of the ground-state density. These two theorems allow – at least in principle – the determination of the ground-state density and energy of a N-particle system by searching for the density that minimizes the energy functional. Extracting the classical Coulomb interaction energy, this Hohenberg-Kohn energy functional takes the form

$$E[n(\mathbf{r})] = \int V_{\text{ext}}(\mathbf{r})n(\mathbf{r})d\mathbf{r} + \frac{1}{2}\int\int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' + G[n(\mathbf{r})]\,, \tag{44}$$

where the functional $G[n(\mathbf{r})]$ contains all other contributions. The functional $G[n(\mathbf{r})]$ is universal in the sense that it is independent of the external potential. If we succeed to find the functional $G[n(\mathbf{r})]$ or a good approximation to it, the immediate advantage of DFT is that, instead of dealing with the full many-body wavefunction, $\Psi(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N)$, we can work with

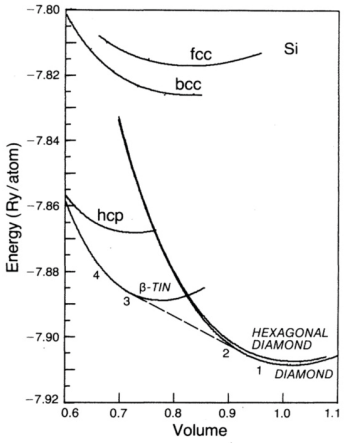


Fig. 4: *The diamond, hexagonal diamond, and β-tin, hcp, bcc, and fcc structural energies (in units of Ry/ atom) as a function of the atomic volume, normalized to the measured free volume for Si. The dashed line is the common tangent of the energy curves for the diamond and the β-tin structures. Results are taken from Ref. [16].*

the much more tractable density, $n(\mathbf{r})$. Although more information is directly accessible from the wavefunction than from the density,

$$n(\mathbf{r}) = \int d\mathbf{r}_2 \ldots \int d\mathbf{r}_N \Psi^*(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N)\Psi(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N) \tag{45}$$

in DFT many physical quantities, like the structural properties or bond strength can be obtained for large systems, where a many-body wavefunction would be impossible to access.
For example, calculations of the ground-state energies for different external potentials, as they result from a variation of the lattice parameters or different crystal structures in a periodic solid, allow the determination of the equilibrium lattice constant, which is nowadays possible to within a few percents. An example is shown in figure 4, which exhibits the calculated total energy of bulk Si calculated for six plausible crystal structures of Si and about ten different lattice constants for each crystal structure fitted to a functional form. The diamond structure is found to be the stablest in agreement with experiment. This calculation also shows that Si will transform to the β-tin structure under high pressure (tangent path 2 – 3). The theoretically determined ground state volume of Si in the diamond structure differs to the experimental results by 1.1%. Also the ground state density obtained for each set of external potential can be used to analyse the bonding behavior of solids. In figure 5 we show the valence electron charge density of a group IV elemental semiconductor (Si) in the diamond structure, and a III-V (GaAs) and a II-VI (ZnSe) compound semiconductor in the zincblende structure as contour plots in the $(1\bar{1}0)$ plane. For Si one observes a charge density that symmetric with respect to the bond center. The charge density contour describes a significant bounding charge between the atoms, which is in agreement (the difference is about 15%) with x-ray scattering data. This charge density represents the covalent bond of Si sp^3 orbitals. As we move from Si to the III-V to the more ionic II-VI compounds, the bond centers are shifted toward the anion sites, the maximum intensities increase, and the bonds become more localized.
Early attempts to use the density as a key parameter for calculations of periodic solids were made by Lenz [18] based on the statistical method of Thomas [19] and Fermi [20]. In this approach, $G[n(\mathbf{r})]$ was considered to contain the kinetic energy density (taken to be proportional to $[n(\mathbf{r})]^{\frac{5}{3}}$). In the Thomas-Fermi-Dirac method $G[n(\mathbf{r})]$ even contains an exchange energy

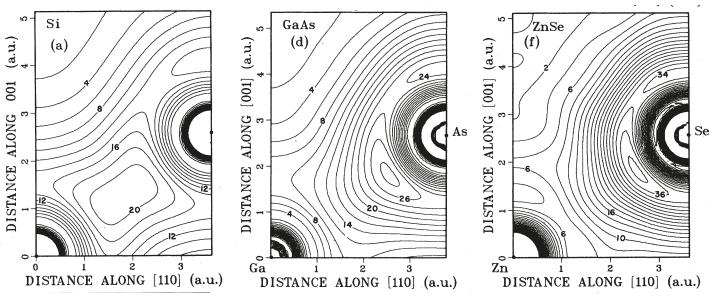


Fig. 5: *Calculated self-consistent valence charge density in a portion of a $(1\bar{1}0)$ plane for Si (left), GaAs (middle) and ZnSe (right) semiconductor. The contours are in units of electrons/unit cell, and the contour interval is two electrons/unit cell. To be consistent with the results of the Ga compounds, the Zn 3d states, which lie more than 5 eV above the bottom of the valence bands with a dispersion of less than 1 eV, were not included. Results are taken from Ref. [17].*

density term proposed by Dirac [21] (proportional to $[n(\mathbf{r})]^{\frac{1}{3}}$). Although the Thomas-Fermi theory has still its applications today, it never became useful as a theoretical method for the prediction of materials properties [22].

The key idea, that made DFT a success, was to extract from $G[n(\mathbf{r})]$ the kinetic energy T_0 of a non-interacting electron system in its ground state, which has the same density distribution, $n(\mathbf{r})$, as the interacting one. In this Kohn-Sham theory [23] a new functional

$$E_{\rm xc}[n(\mathbf{r})] = G[n(\mathbf{r})] - T_0[n(\mathbf{r})] \tag{46}$$

appears, that remains to be determined. $E_{\rm xc}$ is a much smaller term than G and is called exchange-correlation energy functional, since – as we will see below – without $E_{\rm xc}$ our energy functional E would yield just the energy in the Hartree approximation. If we take into account that particle conservation, i.e. $N = \int n(\mathbf{r}) d\mathbf{r}$, has to be ensured, we can formulate the stationarity of E in equation (44) with respect to variations of the ground-state density, n, as

$$\frac{\delta T_0}{\delta n(\mathbf{r})} + V(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + \frac{\delta E_{\rm xc}}{\delta n(\mathbf{r})} - \lambda = 0\,, \tag{47}$$

where the Lagrange parameter λ ensures the particle conservation. Expressing the kinetic energy of the non-interacting particles via their wavefunctions, ϕ_i, we can recast Eq. (47) in the form of an effective single-particle Schrödinger equation, the Kohn-Sham equation:

$$\left[-\frac{1}{2}\nabla^2 + V(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + \frac{\delta E_{\rm xc}}{\delta n(\mathbf{r})} \right] \phi_i(\mathbf{r}) = \varepsilon_i \phi_i(\mathbf{r})\,, \tag{48}$$

which has to be solved self-consistently since $n(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2$. From this point of view, the structure of the Kohn-Sham equations is very similar to the Hartree approach outlined in the last subsection. The index i combines now the k-point, $\mathbf{k}$, and the band index, ν. Note, that without $E_{\rm xc}$ equation (48) reduces to the Hartree equation. Therefore, this last term of the Hamiltonian is called the exchange-correlation potential, often abbreviated as $V_{\rm xc}$, since exchange and correlation are exactly what is missing in the Hartree approximation.

Although λ was introduced as a Lagrange multiplier and also the ε_i's in the Hartree-Fock theory should be strictly be interpreted in this way, it is a common procedure to derive from the ε_i's the bandstructure of a crystal and use the wavefunctions $\phi_i(\mathbf{r})$ as approximations to true quasiparticle wavefunctions. Some justification will be given below and comparison with experimental data often confirms this point of view, but there are also well-known examples, where this interpretation leads to significant "errors", like in the comparison of the bandgaps of semiconductors and insulators with bandstructures derived from these ε_i's.

A second key to the success of DFT was the fact that for the term, which emerged as the exchange-correlation potential in the Kohn-Sham equation (48), numerically simple, but powerful approximations could be found. One of the first interpretation of this term was given by Slater [22] in the context of the Thomas-Fermi method and later in connection with the Hartree-Fock method [11]. Essentially, it describes the aforementioned interaction of a particle with the "hole" that is created by its own presence in the gas of the other particles. This means, that the probability of finding an electron at a position $\mathbf{r}$ reduces the probability of finding another electron at a position $\mathbf{r}'$ nearby, depending of course also on the spin of the two particles (therefore, in the Hartree-Fock method this hole, Eq. (43), has been given the name "exchange hole").

In order to gain some understanding why a rather simple function can be powerful and to obtain some guiding principles in constructing an exchange correlation energy functional, it is useful to write this "hole" (exchange-correlation hole in DFT), $n_{\rm xc}$, in terms of a two-particle correlation function, $g(\mathbf{r},\mathbf{r}')$ [24]:

$$n_{\rm xc}(\mathbf{r},\mathbf{r}') = n(\mathbf{r}')\int_0^1 d\xi\,[g_n(\mathbf{r},\mathbf{r}',\xi)-1] \equiv n(\mathbf{r}')h(\mathbf{r},\mathbf{r}'). \tag{49}$$

Here, $g_n(\mathbf{r},\mathbf{r}',\xi)$ is the correlation function of a system of charged particles where the Coulomb interaction is scaled by a factor $\xi\in[0,1]$, $\xi=0$ defining a noninteracting system and $\xi=1$ the physical one, and a ξ-dependent potential has been added, so that the density, $n(\mathbf{r})$, is kept fixed, i.e. independent of ξ. Additionally, the so called hole function, $h(\mathbf{r},\mathbf{r}')$, was introduced. The exchange correlation energy can then be written as

$$E_{\rm xc}[n(\mathbf{r})] = \frac{1}{2}\int d\mathbf{r}\,n(\mathbf{r})\int d\mathbf{r}'\frac{1}{|\mathbf{r}-\mathbf{r}'|}n_{\rm xc}(\mathbf{r},\mathbf{r}'). \tag{50}$$

Although the exchange-correlation hole can be very complicated in shape, it was soon realized, that only its radial dependence enters in the exchange correlation energy [25]. This means that in practice $E_{\rm xc}$ is rather insensitive to details of shape of $n_{\rm xc}$. Some properties of the exchange-correlation hole can be derived from the definition via the correlation function g. E.g. there is a sum rule, which states that $n_{\rm xc}$ corresponds exactly to one electron, i.e. that

$$\int d\mathbf{r}' n_{\rm xc}(\mathbf{r},\mathbf{r}') = -1 \tag{51}$$

has to be fulfilled. Such relations can guide the construction of exchange-correlation functionals or help to judge the validity of existing approximations to $E_{\rm xc}$.

One of the big surprises in the early days of density functional theory was certainly the fact, that even a simple exchange-correlation functional like the local density approximation (LDA) leads to relatively convincing results. The LDA is in the spirit of the aforementioned Thomas-Fermi-Dirac method and starts from the limit of the homogeneous electron gas, assuming $E_{\rm xc}$

rather as a function than as a functional of $n(\mathbf{r})$. Its success can now be explained by the fact, that the exchange-correlation hole in the local density approximation is of the form

$$n_{\mathrm{xc}}^{\mathrm{LDA}}(\mathbf{r},\mathbf{r}') = n(\mathbf{r}')h_0(|\mathbf{r}-\mathbf{r}'|;n(\mathbf{r}'))\,, \tag{52}$$

where $h_0(|\mathbf{r}-\mathbf{r}'|;n)$ is the hole function of an uniform interacting electron gas of density n. For an uniform density, this exchange-correlation hole satisfies equation (51). For a non-uniform density the sum rule should be at least approximately fulfilled and [26] showed, that in LDA this is on average the case. This, together with the fact that E_{xc} depends only on the spherical average of n_{xc}, is mainly responsible for the success of the LDA.

Also modern, exchange-correlation functionals including gradient corrections are constructed in such a form, that they fulfill certain conditions that are known exactly in different limits (like high or low density, constant or slowly varying density etc.). In this way, exchange-correlation potentials are improved on a parameter-free basis. Alternatively, the functionals (or parts of the functionals, e.g. the correlation energy) can be fitted to numerical results from Quantum Monte Carlo calculations. Another strategy – often used in the chemical literature – is to adjust the functional to yield best results (like bond-length, dissociation energies etc.) for a given set of systems.

Since its first formulation, fifty years ago, DFT became the 'standard model' for the description of electronic structure of solids [27]. Not only the well-defined quantities like density and total energy are taken from DFT to interpret properties of matter, also (as we did in figs. 2 and 3) the eigenvalues ε_i are used to interpret the bandstructure and density of states regularly. At least close to the Fermi level the single-particle states described by DFT (and also other methods that will be discussed below) are in character and energetic order typically well described (see the example of $SrTiO_3$ in Fig. 3). Of course the energetic position is inaccurate, the further away from E_{F}, the more pronounced this deviations get. Band gaps are usually underestimated, e.g. in the shown $SrTiO_3$ by 45%.

The close resemblance between calculated bandstructure and experimental data lead to the fact, that the too small bandgaps in DFT are often called a "DFT problem" but of course, these bandstructures do not describe the electron-removal or electron-addition process that defines the bandgap. Other methods, like the GW approximation to many-body perturbation theory are available for this purpose and will be described at the end of this section.

But we have to be aware that in some cases DFT (i.e. the available approximations to the exchange-correlation energy like LDA) gives an account of the states near the Fermi level that is even qualitatively wrong, e.g. it predicts a metal where in reality an insulator is found. Especially when the electronic structure is very far from the state described by the homogeneous or slowly varying electron gas, this can happen. Typical examples are defects in semiconductors or insulators with atomic-like, isolated states. In figure 6 we show the density of states of an isolated Mn atom that substitutes a Ga atom in GaN. Atomic Mn has the electronic configuration s^2d^5 and at the site of the trivalent Ga we can expect a d^4 occupation of the d states. Indeed, in the DFT calculation we find the d states split in spin and according to their tetrahedral environment. The minority states of Mn are unoccupied. The majority e_g bands are occupied and overlap with the conduction band of GaN while the majority t_{2g} states are in the band gap. There, only two out of three states are filled, the Fermi level lies in the t_{2g} peak. This (unphysical) metallicity leads to the (wrong) prediction of high magnetic ordering temperatures in this dilute magnetic semiconductor [29]. A more realistic description of the strongly localized and correlated Mn d electrons is obtained here when they are 'taken out' of the DFT calculation and

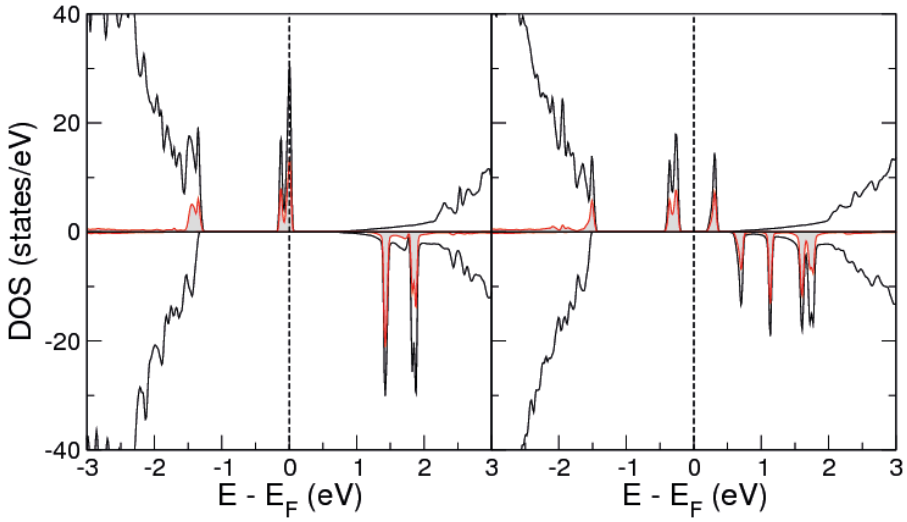


Fig. 6: *Left: LDA density of states of a Mn impurity substituting Ga in a semiconductor (GaN). Positive/negative DOS values indicate majority/minority spin electrons. The Mn states are outlined in red (gray). Right: Result of an LDA+U calculation as described in the text and Ref. [28]: the t_{2g} states are split by almost* 1 *eV and an insulating character is obtained.*

treated in a model that better takes into account their correlated character. This approach, the LDA+U method, will be described in the next section.

3.3 Extensions to DFT: the LDA+U method

Dealing with f and some d transition metals and their compounds it was realized that, while the s, p and some d electrons can successfully be described in standard DFT methods, for the strongly localized electrons a more atomic-like description (e.g. Hartree-Fock) is appropriate [30]. For the same orbital character the degree of electron screening, which is larger for metals than for insulators, is an important parameter deciding on the degree of electron localization. Taking into account the different atomic potentials and the different screening an atomic theory [31] for these localized states can describe the situation quite satisfactorily. Following this approach, Anisimov et al. [32] merged this atomic picture with band theory (i.e. standard DFT), to get a "band approach" to Hubbard-type models: For the localized d and f states, the Coulomb interaction of the electrons is formulated in the spirit of the Anderson model:

$$E_{\rm ee} = \frac{1}{2} U \sum_{i \neq j} n_i n_j \,, \tag{53}$$

where the n's are here the d-orbital occupation numbers, i, j denote the sites of atoms, and U is the famous Hubbard parameter, describing the on-site Coulomb interaction. In the local density

approximation to this model the energy of the $d-d$ interaction is [33]

$$E_{\rm ee}^{\rm LDA} = \frac{1}{2}UN(N-1) \quad \text{where} \quad N = \sum_i n_i \; . \tag{54}$$

If we add E_{ee} from equation (53) to the LDA energy functional, $E_{\rm ee}^{\rm LDA}$ should be subtracted, so that

$$E^{\rm LDA+U} = E^{\rm LDA} + \frac{1}{2}U\sum_{i\neq j} n_i n_j - \frac{1}{2}UN(N-1) \; . \tag{55}$$

This is a simple version of the LDA+U method. Such a modification of the LDA results in a shift of the LDA eigenvalues:

$$\epsilon_i = \frac{dE}{dn_i} = \epsilon_i^{\rm LDA} + U\left(\frac{1}{2} - n_i\right) \tag{56}$$

i.e. more than half-filled bands are shifted down in energy, while less then half-filled bands are shifted up. Despite the formal similarity with the Stoner model, it should be noted that the physical background of this model is quite different [32]. A simple example is given in figure 6, where the LDA+U method was used to correct the positions of the $3d$ states of Mn in a GaN supercell. It is easy to see that the correction has almost no effect on the GaN states, but shifts down the occupied Mn t_{2g} states and pushes the remaining unoccupied levels significantly above the Fermi energy. Thereby, a band gap is opened, its size depends of course on the chosen value of U. Before we turn to the question how to obtain a reasonable estimate for U, we have to refine the model to see, how we can apply the LDA+U method on a certain set of states (e.g. $3d$) at a given atom.

To separate the localized orbitals from the itinerant states, for which the LDA provides already a good description, one chooses a site-centered, $\{l, m\}$ dependent orbital basis, $|\nu, l, m\rangle$, where ν is the site-index of the selected atom and l and m are the angular and azimuthal quantum numbers, respectively. If the density is divided in spin-up ($\alpha \equiv +$) and -down ($\alpha \equiv -$) densities and given by the respective Kohn-Sham orbitals like

$$n^{(+)}(\mathbf{r}) = \sum_i w_i^{(+)} |\phi_i^{(+)}(\mathbf{r})|^2 \quad \text{and} \quad n^{(-)}(\mathbf{r}) = \sum_i w_i^{(-)} |\phi_i^{(-)}(\mathbf{r})|^2 \, , \tag{57}$$

where the weights, $w_i^{(\pm)}$, determine the occupation of the states, we can define a density matrix for spin α in m, m'-space:

$$n_{mm'}^{\alpha\nu} = \sum_i w_i^{\alpha} \langle \nu, l, m | \phi_i^{\alpha} \rangle \langle \phi_i^{\alpha} | \nu, l, m' \rangle . \tag{58}$$

E.g. if we want to apply the LDA+U method on $4f$ states, we need for each spin a 7×7 density matrix, where the diagonal elements give the occupancy of the $l = 3, m = -3, -2, \ldots, 3$ orbitals of the selected atom. Using this density matrix, the electron-electron interaction energy can be formulated as [34]

$$E_{\rm ee} = \frac{1}{2}\sum_{\nu} \sum_{mm'pq}^{\alpha,\beta} n_{mm'}^{\alpha\nu} \left[\langle m, p | V_{\rm ee} | m', q \rangle - \langle m, p | V_{\rm ee} | q, m' \rangle \delta_{\alpha\beta} \right] n_{pq}^{\beta\nu} \tag{59}$$

and used instead of the simpler version presented in Eq. (53). Here, the electron-electron interaction can be expressed in terms of an angular part, contained in a_k, and the radial part that is given by the effective Slater integrals [31], F_k:

$$\langle m,p|V_{\rm ee}|m',q\rangle = \sum_k a_k(m,p,m',q)F_k \quad ; \quad 0\le k\le 2l \tag{60}$$

The Slater integrals F_k can be approximated in terms of the screened Coulomb- and exchange parameters, U and J, e.g. for $l=2$, as

$$U=F_0 \quad ; \quad J=\frac{F_2+F_4}{14} \quad \text{and} \quad \frac{F_4}{F_2}=\frac{5}{8}, \tag{61}$$

and the a_k are sums of integrals of the angular part of the wavefunction with spherical harmonics. Then, we can define an orbital selective potential,

$$V^{\alpha\nu}_{mm'}=\sum_{pq\beta}\left[\langle m,p|V_{\rm ee}|m',q\rangle-\langle m,p|V_{\rm ee}|q,m'\rangle\delta_{\alpha\beta}\right]n^{\beta\nu}_{pq}-\left[U(n^{\nu}-\frac{1}{2})-J(n^{\alpha\nu}-\frac{1}{2})\right]\delta_{mm'}, \tag{62}$$

where $n^{\alpha\nu}=\sum_m n^{\alpha\nu}_{mm}$ and $n^{\nu}=\sum_\alpha n^{\alpha\nu}$. This spin-, site- and l,m-dependent potential enters now the Kohn-Sham equation via

$$\left[-\nabla^2+V^{\alpha}_{LDA}(\vec{r})\right]\phi^{\alpha}_i+\sum_{\nu}\sum_{mm'}V^{\alpha,\nu}_{mm'}\frac{\delta n^{\alpha,\nu}_{mm'}}{\delta\phi^{\alpha}_i}=\epsilon^{\alpha}_i\phi^{\alpha}_i. \tag{63}$$

Thus, we have introduced a Hartree-Fock like potential term that acts on a certain subset of the orbitals, leaving the others (in a first approximation) unchanged. Equation (63) has to be solved self-consistently, until both the density and the density matrix are converged. If the Kohn-Sham equations are solved by expanding the wavefunction into some basis set, for different types of basis sets also a different orbital basis, $|\nu,l,m\rangle$, will be convenient. It is clear, that also the result of the LDA+U calculation will depend to some extent on the choice of the orbital basis, but in practice for the same parameters U and J also qualitatively the same answers are reached. Although the LDA+U method is rather simple and quite successful, it faces the problem that it introduces an external parameter and thus destroys the "*ab initio*" character of the conventional LDA approach. Therefore, concepts to calculate U within constrained DFT [35, 36], in linear response theory [37], or with the GW method [38, 39] (next subsection) have been developed. In the above shown example, figure 6, it can be obtained directly from the total energy difference between a configuration where one electron was added and where one electron was subtracted from the system [28]. This energy difference between $E(N+1)$ and $E(N-1)$ can be calculated from DFT.

Fortunately, in many cases the results do not depend too sensitively on the exact values of U and J. But there are also systems, like $YMnO_3$, where depending on the value of U different magnetic ground-states can be stabilized [40]. A collection of applications of the LDA+U method can be found in reference [33]. For the description of transition metal oxides it is often of fundamental importance to take correlation effects into account, in particular if localized d states are involved. If the transition metal ion is in a d^0 configuration, like Ti in $SrTiO_3$ or TiO_2 (see figure 7), normally the electronic structure is described reasonably well (apart from the well-known underestimation of the band gap in DFT). In the rutile structure the Ti atoms are in the centers of oxygen bipyramids (distorted octahedra) that are connected by corners and edges.

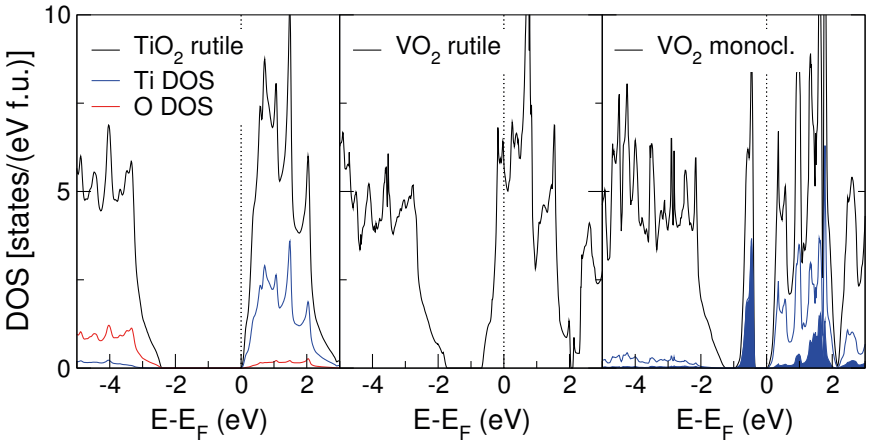


Fig. 7: *Left: Density of states of TiO_2 in the rutile phase. The local O and Ti DOS is shown in red and blue, respectively. The bandgap was corrected using the DFT+U method applied on the O $2p$ and Ti $3d$ states. Middle: DOS of VO_2 in the rutile phase calculated in DFT/GGA. Right: DOS of VO_2 in the monoclinic phase using the DFT+U method. The local Ti DOS is shown in blue, the $d_{x^2-y^2}$ states are marked with the shaded area.*

Like in $SrTiO_3$, the valence band is formed by O $2p$ states, the conduction band by Ti the t_{2g} manifold of the $3d$ states. If we compare TiO_2 to the isostructural VO_2, we realize that the metal ion is now in a d^1 configuration, the Fermi level is now in the V d band (figure 7, middle). This rutile VO_2 phase is indeed metallic and stable above 340 K. Below this temperature VO_2 undergoes a phase transition to the insulating monoclinic (M_1) phase with a dimerization of the V-V pairs of two edge-sharing octahedra. This transition can be seen both as Peierls dimerization and as Mott-Hubbard transition that has to be described with a method that includes both, the correlation aspect (e.g. captured by a Hubbard U) and the sensitivity to non-local interactions (i.e. a k-dependence). Therefore, even advanced methods like the dynamical mean-field theory (DMFT) coupled to DFT (so-called LDA+DMFT) cannot describe this transition in a single-site approximation [41]. Only computationally rather expensive cluster-DMFT studies allow an accurate description [42]. On the other hand, the rather cheap LDA+U approach (figure 7, right) gives already a good impression of the electronic structure of the low-temperature phase. It should be mentioned that the application of the LDA+U method with the same value of U for the rutile phase brings almost no change in the spectrum.

3.4 Quasiparticles and the GW approximation

Up to now, we relied on the concept of single-particle states, which we inherited from the independent-electron approximation. In a many-electron system, the electrons are correlated by the strong Coulomb interaction, the motion of one electron depends on the motion of all other electrons, and it is not at all clear in how far this concept of independent particles is still meaningful. The breakdown of the independent-electron picture questions single-electron concepts like band structure or Fermi surface. Still, in practice these work surprisingly well. In fact, we

can at least retain a nearly-independent-particle picture if we consider quasiparticles instead of electrons (or holes). As we have seen at the beginning of this section, in a system of independent particles the energy to remove a single electron can be determined as the eigenvalue of a single-particle equation like Eq. (34). So we can ask whether it is possible to create an equation similar in structure that yields as an eigenvalue the energy to remove or add a single electron to a many-body system. These energies are the energies of quasiparticles and would be the energies that are typically obtained in experiments like photoemission or inverse photoemission.
It would lead to far to introduce all necessary theoretical concepts to develop an appropriate theory to study this problem. To outline the basic difficulties we will follow here an early paper of many-body perturbation theory [43]. It starts from a single-particle Hamiltonian, e.g. Eq. (37), and assumes that the difference between this Hamiltonian and the true, many-body Hamiltonian can be treated as a perturbation. The energy needed to add a single particle to the N-electron state will differ from the $(N+1)$th eigenvalue of a single-particle Hamiltonian Eq. (37), h_0, by an amount, which is called the self-energy of this particle. A more rigorous derivation shows that a non-local, energy dependent self-energy operator $\Sigma(\mathbf{r},\mathbf{r}',\varepsilon)$ replaces the static exchange correlation potential, $V_{\rm xc}$, in the Kohn-Sham equation (48). The resulting Hamiltonian has the form

$$h_0\phi_i(\mathbf{r}) + \int \Sigma(\mathbf{r},\mathbf{r}',\varepsilon_i)\phi_i(\mathbf{r}')d\mathbf{r}' = \varepsilon_i\phi_i(\mathbf{r})\,. \tag{64}$$

The eigenvalues are now excitation energies, i.e. the energy differences between a N and a $N+1$ particle system (or a N and $N-1$ particle system).
Formally, we can notice a similarity between Eq. (64) and the Hartree-Fock Eq. (40) by defining the following energy independent, e.g. static, self-energy

$$\Sigma^{\rm HF}(\mathbf{r},\mathbf{r}') = \sum_{j,\sigma'} \phi^*_{j,\sigma'}(\mathbf{r}')\phi_{j,\sigma'}(\mathbf{r})\frac{1}{|\mathbf{r}-\mathbf{r}'|} = iG(\mathbf{r},\mathbf{r}',-\eta)v(\mathbf{r},\mathbf{r}')\,. \tag{65}$$

In the last step we wrote the sum over the single-particle states as a Green function with η being an infinitesimally small (positive) time, so that G reduces to the density matrix.
In many-body perturbation theory it turns out that – in a certain approximation – the self energy operator, when Fourier transformed from the energy to the time domain, can be written in a rather similar form:

$$\Sigma(\mathbf{r},\mathbf{r}';\tau) = iG(\mathbf{r},\mathbf{r}',\tau)W(\mathbf{r},\mathbf{r}',\tau+\eta) \tag{66}$$

where G is now the full Green function and W is a screened Coulomb interaction. Generally, $iG(\mathbf{r},\mathbf{r}',\tau)$ describes the probability to measure the presence of an addition particle inserted into a many-body system at a position $\mathbf{r}'$, at a position $\mathbf{r}$ after some time τ. The screened Coulomb interaction W is related to the bare Coulomb interaction $v(\mathbf{r},\mathbf{r}')$ via the dielectric function ϵ,

$$W(\mathbf{r},\mathbf{r}',\varepsilon) = \int \epsilon^{-1}(\mathbf{r},\mathbf{r}'',\varepsilon)v(\mathbf{r},\mathbf{r}'')d\mathbf{r}'' = v(\mathbf{r},\mathbf{r}') + \int n_{\rm ind}(\mathbf{r},\mathbf{r}'',\varepsilon)v(\mathbf{r}',\mathbf{r}'')d\mathbf{r}''\,. \tag{67}$$

Again, we see the effect that the Coulomb potential of the electron repels neighboring charges to give rise to a positive induced charge, $n_{\rm ind}$, that modifies (screens) the bare Coulomb interaction. This behavior reminds of the exchange hole, Eq. (43), of the Hartree-Fock theory or the exchange-correlation hole, Eq. (49), of DFT that gives rise to the exchange-correlation energy, Eq. (50).

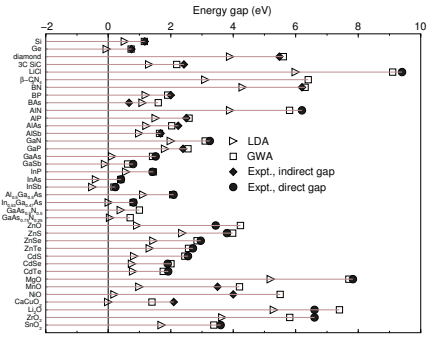


Fig. 8: *Comparison of LDA, GW and experimental band gaps for a variety of materials. Taken from Ref. [46].*

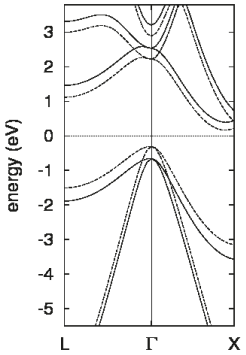


Fig. 9: *LDA band structure (dashed lines) of silicon with GW self-energy corrected valence and conduction bands (solid lines). The GW approximation shifts the corresponding bands up and down, respectively, but leaves the dispersion essentially unaffected.*

It should be noticed that Eq. (66) is a kind of Hartree-Fock (HF) approximation for quasiparticles, while Eq. (65) is the HF approximation for electrons. So, despite the formal similarity to the HF equations we have to keep in mind a couple of important differences: Eq. (64) contains an energy-dependent non-Hermitian self-energy operator. The eigenvalues, ε_i, are complex numbers and the imaginary part leads to a damping term in the time-dependent Schrödinger equation, meaning that the quasiparticles, described by Eq. (64), have a finite lifetime that is proportional to the inverse of the value of the imaginary part. An electron or hole that is added to a many-body system keeps its particle-character for some time, until it "dissipates" into the many-body ensemble.

Hedin [44] provided a set of equations that link all these quantities like the self-energy (containing so-called vertex corrections), the Green function, screened Coulomb interaction and dielectric function. These equations can – in principle – be solved self-consistently. In practice, however, the solution of these equations is far too complicated and commonly an approximation to this equations is solved, which takes Eq. (66) for the self-energy and substitutes the G in this equation by a Green function of the non-interacting system constructed by Kohn-Sham wavefunctions. Also the dielectric function, ϵ, is calculated from these wavefunctions in the random phase approximation. This scheme is commonly termed GW approximation [45] and leads to quite reliable excitations energies, e.g. for bandgaps of semiconductors. Figure 8 shows a comparison of LDA and self-energy corrected band gaps with respective experimental values for a variety of materials. The underestimation within the LDA as well as the improvement by the GW approximation are evident. The principal effect of the GW self-energy correction on the band structure of a semiconductor is to rigidly shift the valence bands up and the conduction bands down, thus opening the band gap. Figure 9 shows this effect for the indirect band gap Si as an example. We find the GW value is close to the experimental value of 1.14 eV at 273 K.

It should be mentioned that this is a method to calculate excitations in a many-body system

where the particle number is changed by one. There are also excitations, which leave the particle number unchanged and are accessible by generalizations of density functional theory, like time-dependent DFT (TDDFT), which provide a way to calculate these types of spectra and are active research fields today [47].

3.5 Short summary: Calculating electronic structure of transition-metal oxides

Before closing this chapter let us summarize the content in three important messages for the calculation of the electronic structure in transition-metal oxides that are often found in resistively switching materials: (i) Mind the subtle interplay between structural properties and electronic structure. As shown above in the example of VO_2, even small structural changes can turn a metal into an insulator. The structure, e.g. tilts and rotations of octahedra in perovskites or other oxides can usually be well described in DFT. (ii) Strong correlations are not well described in DFT in its most widely used approximations. Some modern functionals try to compensate for this, coupling of DFT to models like LDA+U or LDA+DMFT work better but at the expense of an additional parameter that has to be introduced or calculated. Note, that in some cases these correlation effects can be coupled to the structural parameters as well [49]. (iii) Bear in mind that DFT usually underestimates the band gaps. In some cases this is a problem that affects not only the unoccupied states, but has rather profound consequences for the occupied states as well. Hybrid functionals [12] or LDA+U can help if GW calculations become computationally too demanding.

To illustrate the last two points, consider the example of a single oxygen vacancy in $SrTiO_3$: removing a neutral O leaves behind two electrons in the lattice that are localized at the neighboring Ti atoms. Like in VO_2, they have now d^1 configuration and strong correlation effects can be expected [50]. Experimentally, these states are found slightly below the conduction band edge,

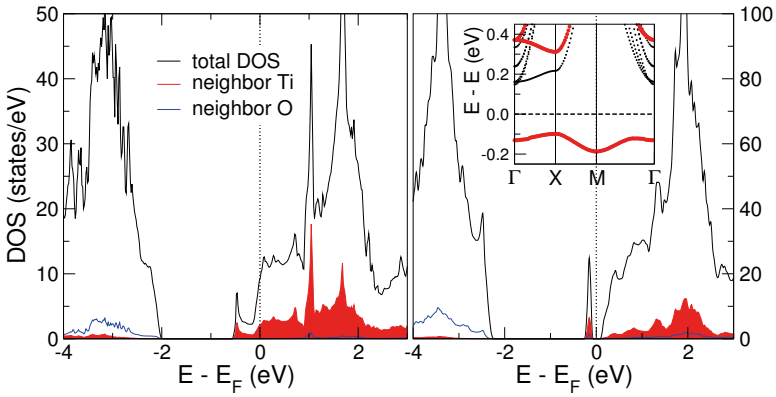


Fig. 10: *Left: Density of states of a* $2 \times 2 \times 2$ *unit cell of* $SrTiO_3$ *containing a single oxygen vacancy: The DFT calculations predicts a metallic ground state. Right: DFT+*U *calculation of the DOS of a oxygen vacancy in* $SrTiO_3$ *using a* $2 \times 2 \times 4$ *unit cell. The defect states are split of and a small band gap of* 0.2 *eV is obtained (adapted from Ref. [48]).*

but well separated to leave the crystal insulating. In a DFT (GGA) calculation these states are already in the conduction band of $SrTiO_3$ and the crystal gets metallic (see figure 10, left). One has to note that experimentally conductive channels have been observed in $SrTiO_3$ and TiO_2 that are induced by oxygen defects and the lateral extension of these states can be as small as 2 nm [51, 52]. Therefore, both the correlated character of the d states and the band gap have to be described reliably to find the experimentally observed behavior (see figure 10, right).

4 Relativistic effects

Why should we care about relativistic effects in solids? We know that these effects become relevant only when velocities are high, close to the speed of light. All electronic wave-packets traveling in the solid are far from that limit. But we have to keep in mind that even conduction electrons spend some time in the vicinity of the atomic nuclei, where the electric fields (potential gradients) can be extremely large and the kinetic energy of these electrons can reach gigantic values there. Of course, these effects are only strong if the atoms forming the lattice are sufficiently heavy, like lead or bismuth (with nuclear numbers $Z = 82$ and 83, respectively). But, as the next example will show, even in lighter atoms some relativistic effects can be observed and should be taken into account when the electronic structure is calculated.

Sb_2Te_3 is a small bandgap semiconductor that crystallizes in hexagonal quintuple layers with a stacking sequence Te-Se-Te-Se-Te. These quintuple layers are only weakly bonded, giving the material a two-dimensional character. A bandstructure of a thin film (6 nm) as calculated in DFT is shown in the left panel of figure 11. Here, we used the so-called scalar-relativistic approximation contains already a few extensions beyond the non-relativistic Schrödinger equation, e.g. the mass-velocity term, but works with two-spinor wavefunctions and the spin enters only via a Zeeman-like term. This approximation was used also in the preceding examples shown in this chapter. On the right of figure 11 we show the same calculation, but now with another relativistic effect, the so-called spin-orbit coupling (SOC), included in the calculation. Two changes are immediately obvious: (i) The band gap is now filled by two linearly dispersing states of opposite spin-direction (as indicated by the red and blue colors) and localized near the surface (indicated by the size of the symbols). (ii) The surface state, marks with red circles in the scalar-relativistic calculation, is now spin-split as can be seen by the blue and red branches in the projected band gap. In this section we will look at the origin of the two SOC-induced features in the above example. SOC effects are not always that prominently visible in other materials, nevertheless it is good to have a feeling what one can expect in a given material and state.

For this, we have to introduce the electrons spin in our considerations. Semi-classically, the electrons "spinning" around its own axis can be thought to be the source of the spin magnetic moment. This should not be confused with the orbital moment, arising from the precessional (orbital) motion of the electron. If we will denote the wavefunction and the spin-label (referred to as spin-up or spin-down) as

$$\psi(\mathbf{r}) = \phi(\mathbf{r})\chi \qquad \text{with} \qquad \chi = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \text{or} \qquad \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{68}$$

we can express the spin, $\mathbf{S}$ as the expectation value of the spin-operator, $\boldsymbol{\sigma}$,

$$\mathbf{S} = \langle\psi|\boldsymbol{\sigma}|\psi\rangle \qquad ; \qquad \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

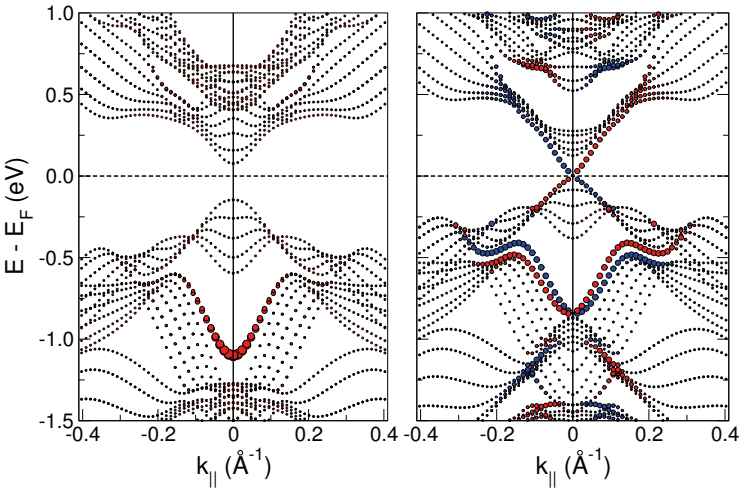


Fig. 11: *Left: Band structure of a 30 layer Sb_2Te_3 film calculated using DFT in the scalar-relativistic approximation. A surface state is marked with red dots, the size of the symbol indicate the surface localization. Right: The same calculation, now with spin-orbit coupling included. The symbol size reflects the spin-polarization of the states, red/blue colors indicate the different spin orientations.*

Of course the Schrödinger equation will provide the wavefunctions ψ, but tells us nothing about the orientation of $\mathbf{S}$. In a collinear case, i.e. when all the spins are oriented along the same direction, for convenience the spins are assumed to be aligned in z-direction.

To give this spin-vector an absolute orientation in space, we first have to introduce a new term in the Hamiltonian that connects the spin-orientation with the axes of the crystal. This is the SOC term mentioned above, which will be discussed on a general basis in the first subsection. In magnetic materials it leads then to a preferential spin orientation in the crystal, however, in most solid state systems, due to chemical bonding, the number of spin-up and spin-down wavefunctions are equal, so that the total spin is zero. Interestingly, even in these spin-compensated systems, that are in total non-magnetic, spin-dependent phenomena can be observed, e.g. in the example shown in figure 11. Due to the fact that we looked in this case at electrons localized on a surface we could observe the so-called Rashba effect, which will be introduced in the second subsection. Finally, we will shortly discuss a certain material class, the topological insulators, that have special metallic states at their surface (filling the gap in figure 11, right) and also rely usually on the presence of spin-orbit coupling.

4.1 Spin-orbit coupling

As a consequence of the Lorentz transformation, an electron that is traveling with a velocity $\mathbf{v}$ on a classical trajectory around the nucleus, experiences an electric field $\mathbf{E}$ (from the potential gradient that arises due to the screened nucleus) as a magnetic field, $\mathbf{B} = \frac{1}{c}(\mathbf{E} \times \mathbf{v})$. This

field will couple to the spin, $\boldsymbol{\sigma}$, of the electron as $-\boldsymbol{\sigma} \cdot \mathbf{B}$.[1] To include this effect on a quantum-mechanical basis, it is necessary to start from relativistic one-electron theory, the Dirac equation. In the Schrödinger equation – even for a magnetic system – there is no term that explicitly includes the spin-operator. But if we include a certain term from the Pauli equation (a two-component approximation to the Dirac equation [54]) we get

$$\left[-\frac{1}{2}\nabla^2 + V(\mathbf{r}) + \frac{\mu_\mathrm{B}}{2c}\boldsymbol{\sigma} \cdot (\mathbf{E}(\mathbf{r}) \times \mathbf{p})\right] \psi_i = \varepsilon_i \psi_i. \tag{69}$$

It is this relativistic correction (factor $\frac{1}{c}$) that leads to the coupling between spin-space ($\boldsymbol{\sigma}$) and lattice ($\mathbf{E}(\mathbf{r})$).
If we assume that the electric field is derived from a spherically symmetric potential, $V(r)$, (as occurs in the vicinity of an atomic nucleus) we can transform this term

$$\boldsymbol{\sigma} \cdot (\mathbf{E}(\mathbf{r}) \times \mathbf{p}) = \boldsymbol{\sigma} \cdot (\nabla V(r) \times \mathbf{p}) = \frac{1}{r}\frac{dV(r)}{dr}\boldsymbol{\sigma} \cdot (\mathbf{r} \times \mathbf{p}) = \frac{1}{r}\frac{dV(r)}{dr}(\boldsymbol{\sigma} \cdot \mathbf{L}) = \xi \boldsymbol{\sigma} \cdot \mathbf{L}, \tag{70}$$

where $\mathbf{L}$ is the orbital momentum operator. This term is called the spin-orbit coupling (SOC) term with the spin-orbit coupling constant ξ. Since the radial derivative of the potential in a crystal will be largest in the vicinity of a nucleus, we can expect that the major contribution to the spin-orbit interaction will come from this region. For an atom ν then r is the radial part of the vector $\mathbf{r}_\nu = \mathbf{r} - \boldsymbol{\tau}_\nu$. Furthermore, since for small r_ν the potential will be Coulomb-like ($V(r) = -\frac{Z}{r}$), its derivative $\frac{\partial V}{\partial r_\nu}$ is proportional to the nuclear number of the atom, Z_ν. We thus expect that ξ will be large for heavy atoms, but small for lighter ones.
Electrons, that are close to the nucleus (i.e. those of the inner shells) will feel the consequences of this spin-orbit coupling most strongly. As it is well known from free atoms, this term will favor the formation of an orbital momentum, $\mathbf{L}$, which is then coupled to the electrons spin. E.g. the p-electrons can form states with a total orbital momentum $L = 1$, coupling then to the electrons spin. We can classify p-states according to their projections on a selected axis (z) by their magnetic quantum numbers $m_z = -1, 0, 1$. Combined with the electrons spin, this will result in a total angular momentum $J = 3/2$ with projections $m_j = 3/2$ or $1/2$. As a consequence of spin-orbit coupling, this results in a level splitting between the $p_{3/2}$ and $p_{1/2}$ states.
In contrast, the valence electrons in a solid will arrange to optimize the chemical bonding, e.g. in a simple cubic lattice p_x, p_y and p_z states will form. The level splitting is then determined by the crystal field. Partially, spin-orbit coupling will interfere and lead to additional level splittings as can be observed e.g. in semiconductors at the center of the Brillouin-zone: In Ge there is a three-fold degenerate state directly below the Fermi-level (figure 12) that splits due to SOC into a doubly degenerate and a singly degenerate one. The former one is closest to the Fermi level in turn consists of two bands with different dispersions, the highly dispersive state is called the light-hole band, the other one is termed heavy hole band. The singly degenerate state at Γ forms the spin-orbit split-off band. In a non-relativistic calculation these bands are degenerate in some high symmetry directions, but when spin-orbit coupling is included a splitting can be observed. As expected, this splitting is smaller in the light Si, but larger in the isoelectronic but heavier α-Sn.

[1] Although this interaction has the form of a Zeeman term (the interaction of the spin with an external magnetic field), its interpretation is not so straightforward: as compared to a classical interpretation, due to kinematical effects a factor of two arises in the expression. The origin of this effect is called Thomas-precession [53].

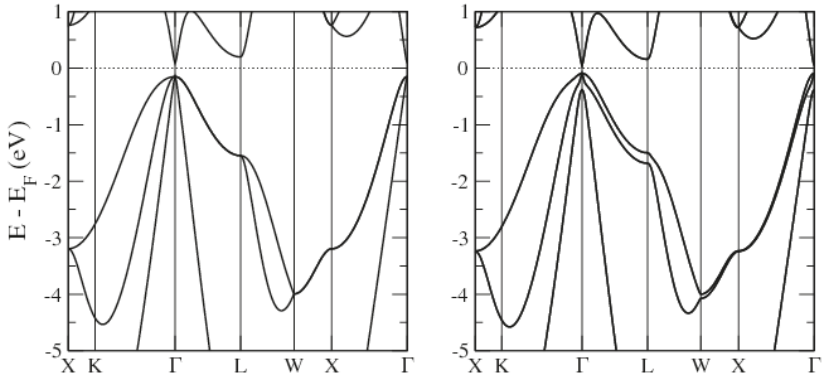


Fig. 12: *Bandstructure of Ge around the Fermi level without spin-orbit coupling (left) and with spin-orbit coupling included (right). Notice, that the three-fold degeneracy of the highest occupied state at the Γ point is split by spin-orbit coupling, as well as the doubly degenerate band along the lines $\overline{\Gamma L}$ and $\overline{\Gamma X}$. The calculation is performed at the experimental lattice constant using the generalized gradient approximation to DFT. Note, that the experimentally observed bandgap of 0.75 eV almost closes in a DFT calculation.*

4.2 The Rashba- and the Dresselhaus effect

In a system without internal or external magnetic field time-reversal symmetry holds, i.e. changing the direction of the arrow of time will not alter the properties of the system. The transformation $t \to -t$ exchanges a particle moving with momentum $\mathbf{k}$ with a particle moving in $-\mathbf{k}$. Time reversal will also invert the precessional motion of the electron and, therefore, its spin. As a consequence, the energy of a right-moving spin-up particle will equal the energy of a left moving spin-down particle,

$$\varepsilon(\mathbf{k},\uparrow) = \varepsilon(-\mathbf{k},\downarrow)\,. \tag{71}$$

In a crystal with inversion symmetry, additionally $\varepsilon(\mathbf{k}) = \varepsilon(-\mathbf{k})$ holds, both for spin-up and spin-down electrons. This means, that the bandstructure is symmetric around the center of the Brillouin-zone, $\mathbf{k} = 0$, and all bands are doubly degenerate. E.g. in the bandstructure in figure 12 shows this degeneracy.
In contrast, crystals without inversion symmetry the degeneracy of the bands can be lifted as a consequence of spin-orbit coupling and only Eq. (71) holds. This can be understood if we realize that a lack of inversion symmetry, $V(\mathbf{r}) \neq V(-\mathbf{r})$, will result in a non-vanishing potential gradient or electric field, $\mathbf{E}(\mathbf{r})$. As we have seen in the last section an electron moving in an electric field will experience this field Lorentz-transformed as $\mathbf{B}$-field and

$$\varepsilon(\mathbf{k},\uparrow) \neq \varepsilon(\mathbf{k},\downarrow)\,. \tag{72}$$

This will, depending on symmetry, result in different consequences for the bandstructures.
Performing a Taylor expansion of the potential $V(\mathbf{r})$, $V(\mathbf{r}) = V_0 + e\mathbf{E}(\mathbf{r})\cdot\mathbf{r} + \cdots$, in lowest order the inversion asymmetry of the potential $V(\mathbf{r})$ is characterized by an electric field $\mathbf{E}(\mathbf{r})$. When electrons with an effective mass m^* propagate with a velocity $\mathbf{v} = \mathrm{d}\varepsilon/\mathrm{d}\mathbf{p} = \frac{1}{m^*}\mathbf{k}$ in an external electric field $\mathbf{E}$ defined in a global frame of reference, then the relativistic Lorentz

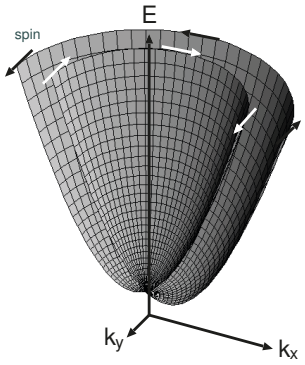


Fig. 13: *Cut through the parabolic energy dispersions of a two-dimensional electron gas in a structure inversion asymmetric (SIA) environment. Indicated are the vector fields of the spin-quantization axes (or the patterns of the spin) at the Fermi surface. As the opposite spins have different energies, the Fermi surface becomes two concentric circles with opposite spins. The effective B-field,* $\mathbf{B}_{\rm eff}$ *is always perpendicular to the propagation direction defined by* $\mathbf{k}_\parallel$.

transformation gives rise to magnetic field $\mathbf{B} = \frac{1}{c}(\mathbf{v} \times \mathbf{E}) = \frac{1}{m^* c}(\mathbf{k} \times \mathbf{E})$ in local frame of the moving electron. The interaction of the spin with this $\mathbf{B}$ field leads then to the so-called Rashba or Bychkov-Rashba Hamiltonian [55, 56]

$$H_{\rm R} = \alpha_{\rm R}\boldsymbol{\sigma}\cdot(\mathbf{p}\times\mathbf{E}) \qquad \text{or} \qquad H_{\rm R} = \alpha_{\rm R}\,\boldsymbol{\sigma}\cdot(\mathbf{k}\times\mathbf{E}) \qquad \text{or} \qquad H_{\rm R} = \alpha_{\rm R}(|\mathbf{E}|)\,\boldsymbol{\sigma}\cdot(\mathbf{k}\times\hat{\mathbf{e}}) \quad (73)$$

describing the Rashba spin-orbit coupling as additional contribution to the kinetic energy. $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ are the Pauli matrices, Eq. (69). The latter two terms are strictly correct only for plane wave eigenstates as, e.g. for a two-dimensional electron gas (2DEG). An important realization of a 2DEGs are electrons in doped semiconductor heterostructures, that support an electron gas at the interface between two materials, e.g. (InGa)As and InP [57]. Another possibility to study the Rashba-effect in 2DEGs was shown in figure 11: on surfaces which support a surface state, e.g. on the Sb_2Te_3 (111) surface [58], the electrons of the surface state move in a potential gradient that is provided by the surface itself (but can also be modified slightly by external electric fields [59]). But also bulk crystals with broken inversion symmetry, like wurzite (ZnO), show this effect and this is also where the first studies by Rashba and Sheka were performed in 1959 [60].

The general features of the Rashba-model can be studied for the 2DEG in a potential with structural inversion asymmetry (SIA) and the corresponding bandstructure are displayed schematically in figure 13. For electrons propagating in the 2DEG extended in the (x, y) plane subject to an electric field normal to the 2DEG, $\hat{\mathbf{e}}_z = (0, 0, 1)$, the Hamiltonian takes the form

$$H = H_{\rm K} + H_{\rm R} = \frac{\mathbf{p}_\parallel^2}{2m^*} + \alpha_{\rm R}\left(\boldsymbol{\sigma} \times \mathbf{p}_\parallel\right)_{|z} = \frac{\mathbf{p}_\parallel^2}{2m^*} + \alpha_{\rm R}\left(\sigma_x p_y - \sigma_y p_x\right), \quad (74)$$

which can be solved analytically. For a Bloch vector in the plane of the 2DEG, $\mathbf{k}_\parallel = (k_x, k_y, 0) = k_\parallel(\cos\varphi, \sin\varphi, 0)$, the eigenstates written as a product of plane wave in space and two-component spinor are

$$\psi_{\pm\mathbf{k}_\parallel}(\mathbf{r}_\parallel) = \frac{e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel}}{2\pi}\frac{1}{\sqrt{2}}\begin{pmatrix} ie^{-i\varphi/2} \\ \pm e^{i\varphi/2} \end{pmatrix} \quad (75)$$

with eigenenergies

$$\varepsilon_\pm(\mathbf{k}_\parallel) = \frac{\mathbf{k}_\parallel^2}{2m^*} + \alpha_{\rm R}\left(\boldsymbol{\sigma} \times \mathbf{k}_\parallel\right) = \frac{\mathbf{k}_\parallel^2}{2m^*} \pm \alpha_R|\mathbf{k}_\parallel| = \frac{1}{2m^*}(k_\parallel \pm k_{\rm SO})^2 - \Delta_{\rm SO}\,, \quad (76)$$

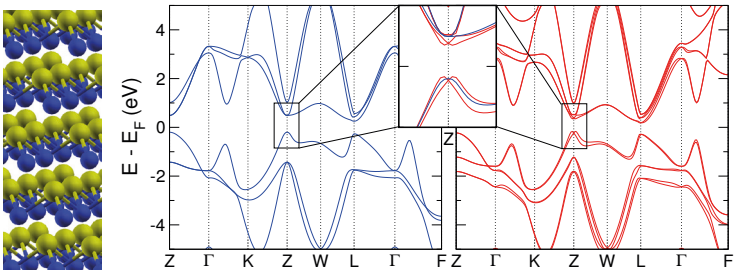


Fig. 14: *Crystal structure of GeTe (left) and band structure calculated without (left) and with (right) spin-orbit coupling included. In particular around the Z-point SOC effects are easy to observe as a splitting of the bands that linear in the momentum (for a magnification, see inset). Blue and yellow spheres indicate Ge and Te atoms, respectively.*

where $\pm$ denotes the spin-up and -down states with respect to a spin orientation axis $\hat{\mathbf{n}}(\mathbf{k}_\parallel)$, local in $\mathbf{k}_\parallel$ space. With the exception of the high-symmetry state $k_\parallel = 0$, we find that the original two-fold degenerate energy paraboloid of the 2DEG in a constant potential is indeed spin-split. This splitting $\varepsilon_+(\mathbf{k}_\parallel) - \varepsilon_-(\mathbf{k}_\parallel) = 2\alpha_\mathrm{R} k_\parallel$ is linear in $k_\parallel$. Due to the presence of the SIA potential and the spin-orbit interaction, the origin of the degenerate parabola is shifted by $k_\mathrm{SO} = m^*\alpha_\mathrm{R}$, but in opposite directions for up- and down-spins with in overall spin-orbit lowering of $\Delta_\mathrm{SO} = m^*\alpha_\mathrm{R}/2$. The orientation axis is given by the expectation value

$$\hat{\mathbf{n}}_\pm(\mathbf{k}_\parallel) = \langle\psi_{\pm\mathbf{k}_\parallel}|\boldsymbol{\sigma}|\psi_{\pm\mathbf{k}_\parallel}|\rangle = \pm\begin{pmatrix}\sin\varphi\\ -\cos\varphi\\ 0\end{pmatrix} \quad \perp \quad \mathbf{k}_\parallel = k_\parallel\begin{pmatrix}\cos\varphi\\ \sin\varphi\\ 0\end{pmatrix}. \qquad (77)$$

We find that the orientation axis is independent of the magnitude $k_\parallel$ and depends only on the direction of the $\mathbf{k}_\parallel$ vector. In fact, it is in the plane of the 2DEG and the orientation axis is perpendicular to the propagation direction of the electron. Considering $\mathbf{k}_\parallel \to -\mathbf{k}_\parallel$, φ changes to $\varphi+\pi$, we find that the spin orientation axis reverses as indicated in figure 13. Thus for $\mathbf{k}_\parallel$ and $-\mathbf{k}_\parallel$ the spin-up and -down states refer to opposite orientations. Defining a global quantization axis along the line $(-\mathbf{k}_\parallel, \mathbf{k}_\parallel)$, e.g. according to $\hat{\mathbf{n}}_\pm(+\mathbf{k}_\parallel)$, then a spin-up state appears as spin-down state if $\mathbf{k}_\parallel$ changes sign. Together with the eigenvalue spectrum given in equation (76) the Kramer degeneracy $\varepsilon_\uparrow(\mathbf{k}_\parallel) = \varepsilon_\downarrow(-\mathbf{k}_\parallel)$ holds. In all, the magnetic moment is zero when averaged over all states $\mathbf{k}_\parallel$. This is consistent with the absence of an $\mathbf{B}$ field.

As an example, where the Rashba effect occurs in a bulk system, we consider here GeTe which, together with Sb_2Te_3, in one of the parent compounds of many phase change materials [61]. In first approximation, GeTe crystallizes in a cubic rock-salt structure. In this arrangement the atoms form alternating planes in (111) direction that consist purely of Ge or Te. In the ground state structure a dimerization along this (111) axis happens and double-layers of Ge and Te are formed (see figure 14, left). This structure is actually very similar to the one of the heavier group V semimetals (structure type A7), which is not surprising if we imagine that Ge (group IV) and Te (group VI) have on average the same number of valence electrons like Sb or Bi. But, in contrast to the A7 structure, the inversion symmetry is broken in the GeTe lattice.

If we look at the bandstructure without SOC effects in figure 14, we recognize that Ge and Te

p-bands form groups of bonding and anti-bonding states. Each band is doubly degenerate. Including spin-orbit coupling, however leads to a lifting of this degeneracy and the mechanism is very similar to the one described above. Of course we have to keep in mind that the states that are affected here are of p character that can, as discussed in the example of Ge in the diamond lattice, form orbital moment carrying states and this complicates the situation as compared to s-like states. But at the conduction and valence band edge it is clearly visible that the splitting is linear in the momentum. This splitting is a function of the polarization of the lattice and disappears in the rock-salt structure. Also proposals to exploit the coupling of polarization and spin-splitting have been put forward for this material [62]: In figure 14, left, each bilayer consists of Te on the upper and Ge on the lower side. If the bonds rearrange (e.g. under the influence of an electric field) so that Ge is on the upper and Te on the lower side, the polarization of the material inverts and also the electric gradients. This, in turn, also inverts the spin-directions of the spin-split bands in GeTe.
That spin-orbit coupling may have important consequences for the one-electron energy levels in bulk semiconductors was first emphasized by Dresselhaus *et al.* [63] already in 1955. Unlike the diamond structure of Si and Ge, the zinc blende structure, in which for example the III-V semiconductor crystallize, exhibit a bulk inversion asymmetry (BIA), i.e. this crystal structure lacks a center of inversion, so that we can have a spin splitting of the electron and hole states at nonzero wave vectors $\mathbf{k}$ as for the Rashba effect even if $\mathbf{B} = 0$. Today, this is called the Dresselhaus effect. The corresponding Dresselhaus Hamiltonian

$$H_{\mathrm{D}} = \alpha_{\mathrm{D}} \left[\sigma_x p_x \left(p_y^2 - p_z^2\right) + \sigma_y p_y \left(p_z^2 - p_x^2\right) + \sigma_z p_z \left(p_x^2 - p_y^2\right)\right] \tag{78}$$

describes the BIA spin splitting due to the Dresselhaus spin-orbit coupling, which produces spin vector fields quite different from those produced by the SIA splitting. One difference is obviously that the Dresselhaus term produced a spin splitting which is proportional to k^3, $\varepsilon_{\mathrm{D}} \propto k^3$, while the spin splitting of the Rashba-term is linear in k, $\varepsilon_{\mathrm{R}} \propto k$. One important difference for the resulting band structures is that the Rashba-term changes the band extremum of a parabolic state from a point to a ring of extrema, while the Dresselhaus term preserves the point character of the band extremum.

4.3 Topological insulators

Let us finally turn to the second SOC-induced feature we observed in the electronic structure of the Sb_2Te_3 film (figure 11), the appearance of linear dispersing states that connect valence- and conduction-band and close the band gap. To understand this observation, we first have a look at the bulk band structure of Sb_2Te_3 without and with spin-orbit coupling included in the calculation (shown in figure 15 left and right, respectively). In both cases we see a band gap, 100 meV without and 150 meV with SOC included. We can tune the spin-orbit coupling strength continuously between zero and its natural value by introducing a scaling factor, λ, that replaces ξ in equation (70) by $\lambda\xi$. Varying now λ between zero and one, we find a closing and reopening of the gap at about $\lambda = 0.5$ (middle of figure 15). This so-called band inversion can also be seen from the symmetry properties of the bands at the Γ-point, here the parity of the wavefunction that is positive if an inversion operation, $\mathcal{I}$, does not change the sign of the wavefunction, $\mathcal{I}\Psi(\mathbf{r}) = \Psi(-\mathbf{r}) = \Psi(\mathbf{r})$, while it is negative if $\mathcal{I}\Psi(\mathbf{r}) = -\Psi(\mathbf{r})$.
In a Gedankenexperiment, imagine you create an interface between Sb_2Te_3 with $\lambda = 1.0$ and Sb_2Te_3 with $\lambda = 0.0$. Although both materials are insulators, the wavefunctions have to be matched from one side to the other and this is not possible without closing the band gap. Due

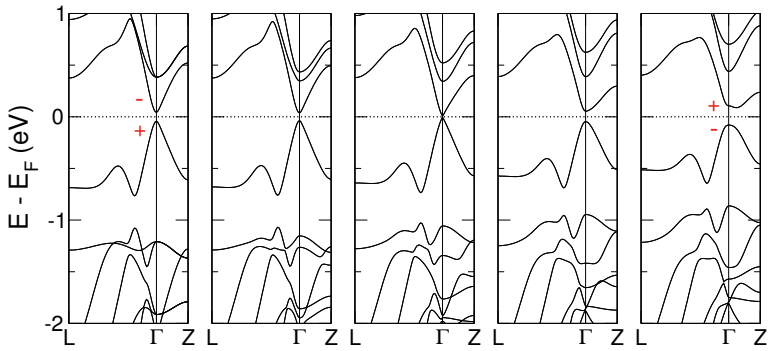


Fig. 15: *Bulk bandstructure of Sb_2Te_3 between the L-point [$\mathbf{k} = (0, \frac{1}{2}, 0)$], the center of the Brillouin zone, Γ, and the Z-point [$\mathbf{k} = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$]. The spin-orbit coupling strength is scaled by $\lambda = 0.0, 0.2, 0.5, 0.8$ and 1.0 from left to right. The parity of the wavefunction is also indicated $(+, -)$ in the outermost plots.*

to the symmetry properties of the two systems at the interface a metallic state appears that is different from interface states that are sometimes induced by broken bonds or local potential variations. In contrast to these "natural" interface states, that can be possibly removed by changing the chemistry at the interface, the symmetry-induced interface states are insensitive to external influences in the interface region because they are consequences of the bulk materials that form the interface. In reality, of course, Sb_2Te_3 does not exist without SOC. But a very similar compound, Sb_2Se_3, exists where the heavier Te ($Z = 52$) is replaced by the lighter Se ($Z = 34$) and SOC effects are weaker. From the point of view of the symmetry properties of the band structure, Sb_2Se_3 is similar to the theoretical model of Sb_2Te_3 with vanishing SOC strength.

Mathematically, the symmetry properties discussed above on the example of Sb_2Te_3 can be shown to be consequences of a certain topology of the Hamiltonian describing a specific material. If two insulating systems are brought in contact and the underlying Hamiltonians have different topology then metallic states have to appear on that boundary. Sb_2Se_3, for example, has the same topology as vacuum and is called topologically trivial. Sb_2Te_3, on the other hand, differs in topology and is called a topological insulator [64]. At the surface of a topological insulator the topology of the Hamiltonian changes and a metallic surface state crosses the gap (see figure 11, right). Without SOC, the topology of Sb_2Te_3 and vacuum are the same and the gap is open (figure 11, left).

It would lead too far to discuss all the fascinating physics of topological insulators but excellent reviews are available, e.g. Ref. [65]. Here, we limit ourselves to some aspects relevant for phase change materials, in particular $Ge_2Sb_2Te_5$ (GST-225). In its hexagonal form, this compound can be seen as combination of quintuple layers of Sb_2Te_3 and two (111) oriented bilayers of GeTe (cf. figure 14). In the energetically favorable Kooi-De Hosson (KH) phase the stacking sequence is Te-Sb-Te-Ge-Te-Ge-Te-Sb-Te-, i.e. the GeTe layers are inserted into the Sb_2Te_3 units, in the metastable Petrov phase the stacking is Te-Ge-Te-Sb-Te-Sb-Te-Ge-Te-, i.e. the GeTe layers are between the Sb_2Te_3 blocks. Surprisingly, the latter phase is a topological insulator, while the former one is topologically trivial [66].

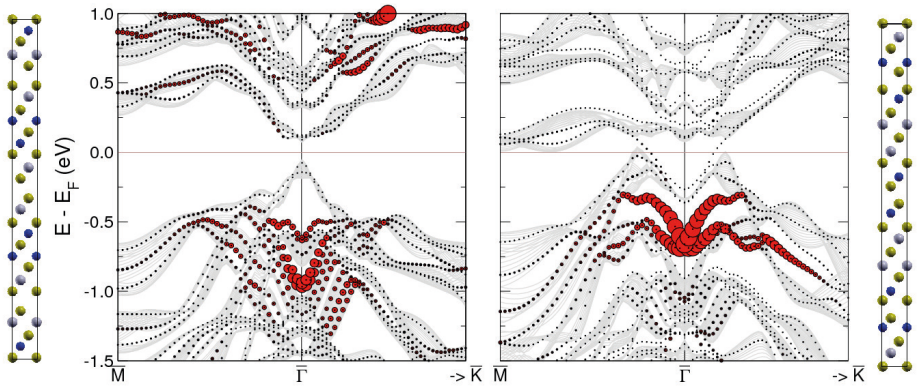


Fig. 16: *Structure and electronic structure of cubic $Ge_2Sb_2Te_5$ in the KH (left) and Petrov (right) stacking sequence. Ge, Sb, and Te atoms are shown in blue, gray and yellow, respectively. The projected bulk band structures in (111) direction are indicated by gray lines in the middle panels, the states of 27 layer thick films are plotted by black/red circles on top. The size of the circles reflects the surface localization of the states.*

Apart from these hexagonal phases, also metastable cubic GST-225 exists where Ge, Te, Sb and vacancies occupy a simple cubic lattice. This cubic form is particularly relevant for applications as phase change material [67]. If the atoms and vacancies order on the cubic (111) planes, stable sequences can again be classified as KH- or Petrov-type. The differences in the electronic structures are significant, as can be seen from figure 16: not only the bulk bandstructure differs (in particular at the Brillouin-zone center), also the surface states differ radically. While the surface of the KH-type cubic GST-225 is still insulating, on the Petrov-type phase surface states appear that close the gap. Also in this form GST-225 is a topological insulator and the existence of this phase was experimentally demonstrated [68]. Clearly, the high structural flexibility of phase change materials and the fact that certain structural motifs lead to the appearance of topological properties (and the conductive boundaries of these phases) have triggered the hope that the transport properties of these materials can be changed by external parameters like ferroelectric polarization [69]. In any case, these examples show that relativistic effects, like spin-orbit coupling, can lead to unexpected modifications of the electronic structure and should not be left out a-priori in the calculation of the band structure even in compounds with just moderately heavy atoms.

References

[1] F. Bloch, Über die Quantenmechanik der Elektronen im Kristallgitter, Z. Physik **52**, 555 (1928).

[2] H. Bethe, Theorie der Beugung von Elektronen in Kristallen, Ann. Physik (Leipzig) **87**, 55 (1928).

[3] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College, Philadelphia, 1976.

[4] V. L. Moruzzi, J. F. Janak, and A. R. Williams, *Calculated Electronic Properties of Metals*, Pergamon, New York, 1978.

[5] M. I. Aroyo, J. M. Perez-Mato, C. Capillas, E. Kroumova, S. Ivantchev, G. Madariaga, A. Kirov, and H. Wondratschek, Bilbao Crystallographic Server I: Databases and crystallographic computing programs, Z. f. Kristallogr. **221**, 15–27 (2006).

[6] J. E. Inglesfield and E. W. Plummer, The Physics of Photoemission, in *Angle resolved photoemission: theory and current applications*, edited by S. D. Kevan, volume 74 of *Studies in surface science and catalysis*, pages 15–61, Amsterdam, 1992, Elsevier.

[7] V. Staemmler, Introduction to Hartree-Fock and CI Methods, in *Computational Nanoscience: Do It Yourself!*, edited by J. Grotendorst, S. Blügel, and D. Marx, volume 31 of *NIC Series*, Jülich, 2006, Research Center Jülich, also available at `http://www.fz-juelich.de/nic-series/volume31/`.

[8] M. Betzinger, Hartree-Fock and quantum chemical correlation methods, in *Computing Solids: Models, ab-initio methods and supercomputing*, edited by S. Blügel, N. Helbig, V. Meden, and D. Wortmann, volume 74 of *Key technologies*, page A3, Jülich, 2014, Research Center Jülich.

[9] L. D. Landau, Theory of the Fermi liquid, Soviet Phys.-JETP **3**, 920 (1956).

[10] J. C. Slater, Note on Hartree's method, Phys. Rev. **35**, 210–211 (1929).

[11] J. C. Slater, A Simplification of the Hartree-Fock Method, Phys. Rev. **81**, 385–390 (1951).

[12] J. Carrasco, F. Illas, N. Lopez, E. A. Kotomin, Y. F. Zhukovskii, S. Piskunov, J. Maier, and K. Hermansson, First principles simulations of F centers in cubic $SrTiO_3$, phys. stat. sol. (c) **2**, 153 (2005).

[13] A. Görling, Exact treatment of exchange in Kohn-Sham band-structure schemes, Phys. Rev. B **53**, 7024 (1996).

[14] C. Franchini, Hybrid functionals applied to perovskites, Journal of Physics: Condensed Matter **26**, 253202 (2014).

[15] P. Hohenberg and W. Kohn, Inhomogeneous Electron Gas, Phys. Rev. **136**, B864–B871 (1964).

[16] M. T. Yin and M. L. Cohen, Microscopic Theory of the Phase Transformation and Lattice Dynamics of Si, Phys. Rev. Lett. **45**, 1004–1007 (1980).

[17] C. S. Wang and B. M. Klein, First-principles electronic structure of Si, Ge, GaP, GaAs, ZnS, and ZnSe. I. Self-consistent energy bands, charge densities, and effective masses, Phys. Rev. B. **24**, 3393–3416 (1981).

[18] W. Lenz, Über die Anwendbarkeit der statistischen Methode auf Ionengitter, Z. Physik **77**, 713–721 (1932).

[19] L. H. Thomas, The calculation of atomic fields, Proc. Cambridge Philos. Soc. **23**, 542–548 (1927).

[20] E. Fermi, Eine statistische Methode zur Bestimmung einiger Eigenschaften des Atoms und ihre Anwendung auf die Theorie des periodischen Systems der Elemente, Z. Physik **48**, 73–79 (1928).

[21] P. A. M. Dirac, Note on Exchange Phenomena in the Thomas–Fermi Atom, Proc. Cambridge Philos. Soc. **26**, 376–385 (1930).

[22] J. C. Slater and H. M. Krutter, The Thomas-Fermi Method for Metals, Phys. Rev. **47**, 559–568 (1934).

[23] W. Kohn and L. J. Sham, Self-Consistent Equations Including Exchange and Correlation Effects, Phys. Rev. **140**, A1133–A1138 (1965).

[24] W. Kohn and P. Vashista, General density functional theory, in *Theory of the Inhomogeneous Electron Gas*, edited by S. Lundqvist and N. H. March, pages 79–147, New York, 1983, Plenum.

[25] O. Gunnarsson, M. Jonson, and B. I. Lundqvist, Descriptions of exchange and correlation effects in inhomogeneous electron systems, Phys. Rev. B **20**, 3136–3164 (1979).

[26] O. Gunnarsson and B. I. Lundqvist, Exchange and correlation in atoms, molecules, and solids by the spin-density-functional formalism, Phys. Rev. B **13**, 4274–4298 (1976).

[27] R. O. Jones, Density functional theory: Its origins, rise to prominence, and future, Rev. Mod. Phys. **87**, 897 (2015).

[28] T. Fukushima, H. Katayama-Yoshida, K. Sato, G. Bihlmayer, P. Mavropoulos, D. S. G. Bauer, R. Zeller, and P. H. Dederichs, Hubbard U calculations for gap states in dilute magnetic semiconductors, J. Phys.: Cond. Matter **26**, 274202 (2014).

[29] K. Sato, L. Bergqvist, J. Kudrnovský, P. H. Dederichs, O. Eriksson, I. Turek, B. Sanyal, G. Bouzerar, H. Katayama-Yoshida, V. A. Dinh, T. Fukushima, H. Kizaki, and R. Zeller, First-principles theory of dilute magnetic semiconductors, Rev. Mod. Phys. **82**, 1633 (2010).

[30] D. van der Marel and G. A. Sawatzky, Electron-electron interaction and localization in d and f transition metals, Phys. Rev. B **37**, 10674–10684 (1988).

[31] J. C. Slater, The theory of complex spectra, Phys. Rev. **34**, 1293–1322 (1929).

[32] V. I. Anisimov, J. Zaanen, and O. K. Andersen, Band theory and Mott insulators: Hubbard U instead of Stoner I, Phys. Rev. B **44**, 943–954 (1991).

[33] V. I. Anisimov, F. Aryasetiawan, and A. I. Lichtenstein, First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LSDA+U method, J. Phys.: Condens. Matter **9**, 767–808 (1997).

[34] A. I. Liechtenstein, V. I. Anisimov, and J. Zaanen, Density-functional theory and strong interactions: Orbital ordering in Mott-Hubbard insulators, Phys. Rev. B **52**, R5467–R5470 (1995).

[35] P. H. Dederichs, S. Blügel, R. Zeller, and H. Akai, Ground States of Constrained Systems: Application to Cerium Impurities, Phys. Rev. Lett. **53**, 2512–2515 (1984).

[36] I. V. Solovyev, P. H. Dederichs, and V. I. Anisimov, Corrected atomic limit in the local-density approximation and the electronic structure of d impurities in Rb, Phys. Rev. B **50**, 16861–16871 (1994).

[37] M. Cococcioni and S. de Gironcoli, Linear response approach to the calculation of the effective interaction parameters in the LDA+U method, Phys. Rev. B **71**, 035105 (2005).

[38] F. Aryasetiawan, M. Imada, A. Georges, G. Kotliar, S. Biermann, and A. Lichtenstein, Frequency-dependent local interactions and low-energy effective models from electronic structure calculations, Phys. Rev. B **70**, 195104/1–8 (2004).

[39] I. V. Solovyev and M. Imada, Screening of Coulomb interactions in transition metals, Phys. Rev. B **71**, 045103/1–11 (2005).

[40] S. Picozzi, K. Yamauchi, G. Bihlmayer, and S. Blügel, First-principles stabilization of an unconventional collinear magnetic ordering in distorted manganites, Phys. Rev. B **74**, 094402 (2006).

[41] A. Liebsch, H. Ishida, and G. Bihlmayer, Coulomb correlations and orbital polarization in the metal-insulator transition of $\mathrm{VO_2}$, Phys. Rev. B **71**, 085109.

[42] S. Biermann, A. Poteryaev, A. I. Lichtenstein, and A. Georges, Dynamical Singlets and Correlation-Assisted Peierls Transition in $\mathrm{VO_2}$, Phys. Rev. Lett. **94**, 026404 (2005).

[43] J. G. W. Pratt, Generalization of Band Theory to Include Self-Energy Corrections, Phys. Rev. **118**, 462 (1959).

[44] L. Hedin, New Method for Calculating the One-Particle Green's Function with Application to the Electron-Gas Problem, Phys. Rev. **139**, A796–A823 (1965).

[45] F. Aryasetiawan and O. Gunnarsson, The GW method, Rep. Prog. Phys. **61**, 237–312 (1998).

[46] W. G. Aulbur, L. Jönsson, and J. W. Wilkins, Quasiparticle calculations in solids, in *Solid State Physics*, edited by H. Ehrenreich and F. Spaepen, volume 54, pages 1–218, New York, 2000, Academic Press.

[47] G. Onida, L. Reining, and A. Rubio, Electronic excitations: density-functional versus many-body Green's-function approaches, Rev. Mod. Phys. **74**, 601–659 (2002).

[48] K. Szot, G. Bihlmayer, and W. Speier, Nature of the Resistive Switching Phenomena in TiO_2 and $SrTiO_3$: Origin of the Reversible Insulator-Metal Transition, Solid State Physics **65**, 353 (2014).

[49] E. Pavarini, E. Koch, and A. I. Lichtenstein, Mechanism for Orbital Ordering in $KCuF_3$, Phys. Rev. Lett. **101**, 266405 (2008).

[50] C. Lin and A. A. Demkov, Electron Correlation in Oxygen Vacancy in $\mathrm{SrTiO_3}$, Phys. Rev. Lett. **111**, 217601 (2013).

[51] K. Szot, G. Bihlmayer, W. Speier, and R. Waser, Switching the electrical resistance of individual dislocations in single crystalline $SrTiO_3$, Nature Mater. **5**, 312 (2006).

[52] M. Rogala, G. Bihlmayer, W. Speier, Z. Klusek, C. Rodenbücher, and K. Szot, Resistive switching of a quasi-homogeneous distribution of filaments generated heat-treated TiO_2 (110)-surfaces, Advanced Functional Materials **25**, 6382 (2015).

[53] J. D. Jackson, *Classical electrodynamics*, Wiley & Sons, 1962.

[54] H. A. Bethe and E. E. Salpeter, *Quantum Mechanics of One- and Two-Electron Systems*, Plenum, New York, 1977.

[55] Y. A. Bychkov and E. I. Rashba, Oscillatory effects and the magnetic-susceptibility of carriers in inversion-layers, J. Phys. C: Solid State Phys. **17**, 6039 (1984).

[56] Y. A. Bychkov and E. I. Rashba, Properties of a 2D eelctron-gas with lifted spectral degeneracy, Sov. Phys. JETP Lett **39**, 78 (1984).

[57] T. Schäpers, J. Knobbe, and V. A. Guzenko, Effect of Rashba spin-orbit coupling on magnetotransport in InGaAs/InP quantum wire structures, Phys. Rev. B **69**, 235323 (2004).

[58] C. Pauly, G. Bihlmayer, M. Liebmann, M. Grob, A. Georgi, D. Subramaniam, M. R. Scholz, J. Sànchez-Barriga, A. Varykhalov, S. Blügel, O. Rader, and M. Morgenstern, Probing two topological surface bands of Sb_2Te_3 by spin-polarized photoemission spectroscopy, Phys. Rev. B **86**, 235106 (2012).

[59] G. Bihlmayer, Y. M. Koroteev, P. M. Echenique, E. V. Chulkov, and S. Blügel, The Rashba-effect at metallic surfaces, Surf. Sci. **600**, 3888 (2006).

[60] E. I. Rashba and V. I. Sheka, Symmetry of Energy Bands in Crystals of Wurtzite Type: II. Symmetry of Bands Including Spin-Orbit Interaction, Fiz. Tverd. Tela: Collected Papers **2**, 162–176 (1959) (for an english translation, see supplement of New J. Phys. **17**, 050202 (2015)).

[61] W. Wełnic, A. Pamungkas, R. Detemple, C. Steimer, S. Blügel, and M. Wuttig, Unravelling the interplay of local structure and physical properties in phase-change materials, Nature Materials **5**, 56 – 62 (2006).

[62] D. DiSante, P. Barone, R. Bertacco, and S. Picozzi, Electric Control of th Gian Rashba Effect in Bulk GeTe, Advanced Materials **25**, 509 – 513 (2013).

[63] G. Dresselhaus, Spin-orbit coupling effects in zinc blende structures, Phys. Rev. **100**, 580 (1955).

[64] L. Fu and C. L. Kane, Topological insulators with inversion symmetry, Phys. Rev. B **76**, 045302 (2007).

[65] Y. Ando, Topological Insulator Materials, J. Phys. Soc. Jpn. **82**, 102001 (2013).

[66] J. Kim, J. Kim, and S.-H. Jhi, Prediction of topological insulating behavior in crystalline Ge-Sb-Te, Phys. Rev. B **82**, 201312 (2010).

[67] J.-B. Park, G.-S. Park, H.-S. Baik, J.-H. Lee, H. Jeong, and K. Kim, Phase-Change Behavior of Stoichiometric $Ge_2Sb_2Te_5$ in Phase-Change Random Access Memory, J. Electrochem. Soc. **154(3)**, H139–H141 (2007).

[68] C. Pauly, M. Liebmann, A. Giussani, J. Kellner, S. Just, J. Sànchez-Barriga, E. Rienks, O. Rader, R. Calarco, G. Bihlmayer, and M. Morgenstern, Evidence for topological band inversion of the phase change material $Ge_2Sb_2Te_5$, Appl. Phys. Lett. **103**, 243109 (2013).

[69] J. Tominaga, A. V. Kolobov, P. Fons, T. Nakano, and S. Murakami, Ferroelectric Order Control of the Dirac-Semimetal Phase in GeTe-Sb_2Te_3 Superlattices, Adv. Mater. Interfaces **1**, 130027 (2014).

A 3 Lattice disorder in ionic crystals

Felix Gunkel
Institute of Electronic Materials, IWE2
RWTH Aachen University

Contents

1 Introduction **2**

2 Fundamental thermodynamic processes **3**
2.1 Concept of minimum Gibbs energy 4
2.2 Defect formation processes and conservation rules 5
2.3 Solid state chemical reactions and law of mass action 6

3 Lattice disorder in crystalline solids **7**
3.1 Types of lattice disorder 8
3.2 Special character of oxides – oxygen exchange with ambient atmosphere 9
3.3 Defect notation 10

4 Electronic disorder **11**
4.1 Intrinsic electronic disorder 11
4.2 Extrinsic electronic disorder 12

5 Kinetic limitations **13**

6 An example: $SrTiO_3$ **14**
6.1 General properties of $SrTiO_3$ 14
6.2 Defect concentrations in thermodynamic equilibrium 15
6.3 Acceptor-doped $SrTiO_3$ 19
6.4 Donor-doped STO 22

7 Grain boundaries and surfaces **25**

8 Extended defect structures **28**

9 Amorphous materials **31**
9.1 Defects in amorphous solids – concept of dangling bonds 32

10 Summary **32**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

In the past decades, semiconductors have been established in cleaner and cleaner constitutions. There has been a continuous competition within the electronics community to suppress as much as possible lattice disorder and defects, in order to optimize the transport properties of the compounds. This development is driven by Moore´s law requiring constantly increasing performance of the established CMOS technologies.

In the search for novel concepts, that may overcome the limits of CMOS-based electronics and information technology, however, one observes a change of paradigm. New concepts e.g. for data storage have been suggested that make use of lattice disorder. [1] Thus, these concepts aim to exploit and to functionalize disorder in a material. Among these concepts, storage of information in the *lattice structure* of a material (phase change memory) and in the *atomic configuration* and *local stoichiometry* of a material (e.g. valence change memory, VCM, and electrochemical metallization cells, ECM) have attracted special attention and represent focus topics of this Spring School.

The prevention of unwanted defect formation and the *control and functionalization of order and disorder* requires a profound knowledge of the defect formation processes. Therefore, the understanding of lattice disorder is very important for the field of nanoelectronics.

For this, multiple disciplines such as physics, solid state chemistry, material science, and electronic engineering have to be combined. As a general definition, physicists would rather term this discipline *thermodynamics* and *lattice disorder*, while electrochemists might favor the term *defect chemistry*.

The interdisciplinary character is further emphasized by the fact that nowadays we are interested not only in the “cleanest” form of solids, this is a single crystal, but also in thin films, in (ploy-)crystalline and in amorphous manner, as well as in ceramics. Defect chemistry thus affects several scientific communities with diverse aims and backgrounds.

The general field of lattice disorder in solids cannot be covered in a single lecture. Therefore, this lecture will focus on the behavior and the properties of *ionic* materials. It will be discussed how crystal defects and lattice disorder affect the physical properties of ionic materials, with emphasis on the electronic properties of complex oxides.

As we will see, complex oxides can undergo a full transition from electronically insulating behavior to metallic conduction depending on the particular (local) defect structure, making these materials suitable e.g. for the realization of VCM-type memrisitive memories.

More specifically, this lecture will concentrate on a certain family of complex oxides, namely *perovskite oxides*, exhibiting a huge variety of properties. Perovskites exist in many chemical compositions, all sharing a similar crystal structure. Therefore, perovskite oxides reflect a fascinating playground for combining various material compositions (and properties). Modern thin film deposition techniques furthermore allow to mix these materials on the nanoscale in form of epitaxial thin films and superlattices, in which not only bulk material properties are combined, but also novel phases may arise at interfaces [2].

Perovskite oxides address a plethora of applications beyond memristive memories and cover the full range of physics. Among the perovskites, one finds high-k dielectrics, thermoelectric materials, (oxygen) ion conductors and catalytic materials (e.g. for water splitting applications), varistors and oxygen-sensing materials. Moreover, the involved physics cover electronic ef-

fects, such as high-mobility electron gases, superconductivity, novel magnetism (super-exchange, double exchange), ferroelectricity, multiferroics, Mott-type metal-insulator transitions and structural phase transitions, and many more.

Interestingly, these diverse physical effects are almost exclusively controlled and/or influenced by the defect structure of the involved materials, reflecting the fundamental importance of understanding lattice disorder and defect formation.

In the remainder of this chapter, we will first discuss the general thermodynamic driving forces that mediate defect formation in the bulk of materials, based on the thermodynamic principle of minimizing Gibbs energy. Afterwards, different types of defects will be introduced. After defining conservation rules that apply for defect formation processes in solid state matter, we will then discuss the effect of extrinsic and intrinsic defects on the electronic properties of a perovskite oxide model system, namely $SrTiO_3$.

While bulk defect concentrations will be deduced in an electro-neutral approach, we will furthermore address the effect of space charges, electric fields and electro-static potentials. These effects are especially important in the surrounding of grain boundaries in poly-crystalline ceramics as well as at surfaces and interfaces, where the defect concentrations can differ drastically from the bulk values. Finally, we address the impact of extended defect structures and compare crystalline and amorphous materials.

A more detailed elaboration of the defect formation processes in solids and the defect chemistry of oxides beyond the scope of this lecture is provided by *Smyth* [3] and *Catlow* [4].

2 Fundamental thermodynamic processes

The study of lattice disorder has a longstanding history in science, based on the work of Frenkel, Schottky, Wagner, Gibbs, and many more. At its heart, it reflects the study of deviations from the ideal structure of a solid and the resulting consequences for the material properties. Most importantly, defects induce electrical resistance in a solid, as they disturb the ideally periodic potential seen by electrons and thus induce scattering and momentum relaxation.

The ideal structure of solid is given by a single crystal, in the sense that a single crystal is the state of minimum potential energy for a solid. It thus forms naturally over time. The crystal lattice is characterized by a periodic pattern of lattice sites on which atoms or ions (or molecules) are arranged. The particular structure of the lattice as well as the particular distance between lattice sites (lattice spacing) are determined by the minimum energy of the configuration.

Any deviation from the perfect periodic lattice is called *defect*. As we will see, defects naturally exist in any real crystal. As stated above, however, any defect will result in an energy increase of the system. So why do defects exist at all? The answer to this question is provided by thermodynamics, more precisely by the *entropy* of a given state.

While the formation of a defect generally costs energy, the formation of a defect comes along with a gain in entropy. Therefore, the proper quantity describing the thermodynamic equilibrium state of real systems is the so-called *Gibbs energy* – which considers not only energetics but also the entropy of a configurational state of matter. Crystal defects then arise from the thermodynamic principle of minimizing Gibbs energy for a given configuration of matter in a solid (see Fig. 1).

2.1 Concept of minimum Gibbs energy

Gibbs energy, G, of a thermodynamic ensemble is defined as

$$G = H - TS\,, \tag{1}$$

where H denotes the enthalpy of the ensemble, S the entropy of the ensemble, and T the temperature of the ensemble. Both $H=H_n$ and $S=S_n$ depend on the thermodynamic state n of the ensemble. The negative sign in eq. (1) indicates that at finite temperature a gain in entropy may result in a decrease of Gibbs energy for the ensemble.

S_n is proportional to the logarithm of the probability, p_n, for the ensemble to be in the particular thermodynamic state n among all possible states of the ensemble. For a solid, S_n is then essentially given by the number of non-equivalent ways, Ω_n, to arrange atoms (or ions or molecules) on the available lattices sites of the crystal

$$S_n = k_B \ln \Omega_n\,, \tag{2}$$

The proportionality factor k_B is called Boltzmann´s constant.

Starting from an ideal crystal structure with N lattice sites (per volume), there is only one non-equivalent way to arrange N atoms on the N lattice sites. In this case one gets $\Omega_0=1$, and thus $S_0=0$. H_0 corresponds to the crystal formation energy, which we may choose as zero in our energy scale.

If we now consider n vacancy defects – i.e. instead of N atoms we arrange $N'=N-n$ on N lattice sites – we have

$$\Omega_n = \frac{N!}{(N-N')!N'!} = \frac{N!}{n!(N-n)!}\,. \tag{3}$$

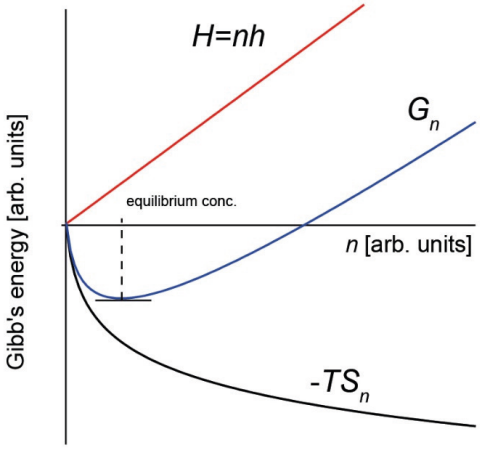


Fig. 1: *Schematic of Gibbs energy as a function of vacancy concentration n. While H=nh increases linearly with n, -TS follows a logarithmic dependence on n. As a result, the minimum Gibbs energy of the ensemble is achieved at finite n. This minimum in Gibbs energy defines the equilibrium defect concentration established at certain temperature.*

For the formation of each single vacancy defect, one has to pay an energy h in order to break bonds and to remove the entity from the crystal. As a result, one gets Gibbs energy

$$G_n = nh - k_B T \ln \frac{N!}{n!(N-n)!}. \tag{4}$$

n now represents the number of vacancy defects incorporated in a crystal at a given temperature T. h is the required defect formation enthalpy involved in a single defect formation process.

Considering a dilute system, i.e. $n \ll N$, the entropy term can be simplified using the Stirling approximation

$$\ln \frac{N!}{n!(N-n)!} \approx \left[N \ln N\right] - \left[n \ln n\right] - \left[(N-n)\ln(N-n)\right]. \tag{5}$$

Using this simplified expression, one can easily calculate the equilibrium defect concentration of minimum Gibbs energy by minimizing eq. (4) with respect to n

$$\begin{aligned} \frac{dG}{dn} &= h - k_B T\left(-\ln(n) - 1 + \ln(N-n) + 1\right) \\ &= h - k_B T \ln\left(\frac{N-n}{n}\right) \approx h - k_B T \ln\left(\frac{N}{n}\right) \\ &= 0. \end{aligned} \tag{6}$$

As a result, we finally arrive at

$$n = N \exp\left(-\frac{h}{k_B T}\right). \tag{7}$$

Equation (7) thus describes the concentration of vacancy defects naturally incorporated in thermodynamic equilibrium into a solid a given temperature T.

Analyzing eq. (7) in detail, we see that at zero temperature $n \to 0$, while $n \to N$ for infinite temperatures. Thus, the equilibrium concentration increases exponentially with increasing temperature because of the increasing entropy contribution to Gibbs energy of the system. The ideal crystal with $n = 0$ exists in equilibrium only at zero absolute temperature.

The main parameter determining the actual defect concentration is the formation enthalpy h. Hence, defect species with low formation enthalpy will exist in much higher density than defect species with a larger formation enthalpy.

2.2 Defect formation processes and conservation rules

Defect formation processes in solids are not independent. In most cases, the formation of a certain defect implies the formation of other compensating defects. As we will see, this is particularly the case for charged defects, as the charge of the formed defect has to be compensated by a counter charge in order to guarantee charge neutrality. Besides charge neutrality, there is a number of conservation rules that have to be applied in a solid state defect formation process. This interdependence is typically formulated in terms of chemical reaction equations obeying the required conservation rules as listed below

1. ***Conservation of mass***

   Atoms can be neither created nor destroyed within a closed system.

2. ***Conservation of charge***

   The bulk of a solid crystal is charge neutral. Therefore, charged defects must be formed in combinations that are overall charge neutral.

3. ***Conservation of structure*** (lattice site ratios):

   The generation of lattice disorder must not violate the inherent ratio of lattice sites in the structure. For instance, in a rock-salt structure such as AgCl the ratio of cation sites (e.g. Ag^+) and anion sites (e.g. Cl^-) must be 1:1.

4. ***Conservation of electronic states*** (band structure):

   The total number of electronic states in a system derives directly from the electronic states of the component atoms and must be conserved.

As we will see, conservation rule 2 (charge neutrality) can be violated locally in space charge regions, such as at interfaces, surfaces and grain boundaries. In these cases, local charge neutrality is replaced by global charge neutrality. Note, however, that the systematic and rigorous application of the conservation rules listed above is the very basic concept of lattice disorder and defect chemistry models.

2.3 Solid state chemical reactions and law of mass action

Defect formation process are often expressed in terms of chemical reaction equations

$$\sum_i a_i \mathrm{A}_i \rightleftharpoons \sum_j a_j \mathrm{A}_j \,, \tag{8}$$

where A_i denote the reactants involved in the reaction and A_j denote the products of the chemical reaction. $a_{i,j}$ represent the corresponding concentration coefficients for reactants (i) and products (j). Solid state chemical reactions must meet the requirements of the conservation rules listed in the forgoing section.

Chemical reactions can generally run into both directions, i.e. reactants → product and products → reactants, simultaneously. In thermodynamic equilibrium, the reaction rates for both directions are equal. Thus, the equilibrium concentrations of all products and reactants are given by constant equilibrium values.

Similar to the derivation of eq. (7), one can derive these equilibrium concentrations from the principle of minimizing Gibbs free energy. However, as the formation of a certain species now is coupled to the formation or removal of other species, we have to minimize the total Gibbs energy of the ensemble

$$G = \sum_i G_i(n_i) + \sum_j G_j(n_j). \tag{9}$$

$n_{i,j} = [\mathrm{A}_{i,j}]$ denotes the concentrations of the species i (reactant) and j (products). For minimization of eq. (9), one has to keep in mind that all $n_{i,j}$ are now coupled via the chemical reaction equation (8). Thus, when varying the concentration of species k, the concentrations of species k' will vary, too.

In particular, we have $dn_k = a_k a_{k'}^{-1} dn_{k'}$ for reactants and $dn_k = -a_k a_{k'}^{-1} dn_{k'}$ for products, when varying any reactant concentration $n_{k'}$ by $dn_{k'}$. With these considerations, we can now minimize G with respect to an arbitrarily chosen reactant concentration n_k

$$\frac{dG}{dn_k} = \sum \frac{a_i}{a_k} \frac{\partial G_i}{\partial n_i} - \sum \frac{a_j}{a_k} \frac{\partial G_j}{\partial n_j} = 0 . \tag{10}$$

Here, we defined the *chemical potential* $\eta_k = \partial G / \partial n_k$ as the partial derivative of Gibbs energy with respect to the particle number n_k.

Using eq. (4), one arrives at

$$\frac{dG}{dn_k} = \sum_{\text{react.}} \frac{a_i}{a_k} \left(h_i - \ln \frac{N_i}{n_i} \right) - \sum_{\text{produc.}} \frac{a_j}{a_k} \left(h_j - \ln \frac{N_j}{n_j} \right) = 0 , \tag{11}$$

which is the *law of mass action*. After some calculus, one gets

$$\frac{\prod n_j^{a_j}}{\prod n_i^{a_i}} = K_0 \exp\left(-\frac{\Delta H}{k_B T} \right). \tag{12}$$

Here, K_0 is a (mostly) temperature-independent constant, ΔH is the reaction enthalpy that can be identified with the weighted sum of all single formation enthalpies involved in the chemical reaction

$$\Delta H = \sum a_j h_j - \sum a_i h_i . \tag{13}$$

Equation (12) yields a relation between the different concentrations of chemical entities involved in a chemical reaction und thus reflects the coupling of the concentrations under certain thermodynamic equilibrium conditions, which are – in the first place – determined by the temperature of the system.

In terms of lattice disorder, we will use this concept of chemical reactions and the concept of the law of mass action, in order to derive the equilibrium concentrations of (point) defect concentrations in a solid, which are also coupled among each other via chemical reaction equations.

3 Lattice disorder in crystalline solids

As states before, defects are defined as any deviation from the ideal crystal lattice. This general statement can be refined using the parameter of the dimensionality of a certain defect. The characteristic defect structures in crystals range from (nano-)voids (3-dimensional (3-D) defect), to stacking faults and shear planes (2-D), to dislocations (1-D), to point defects (0-D).

While all defect structures have important implications for the physical and particularly electronic properties of the materials, we will concentrate the discussion of lattice disorder to point-like (0-D) deviations from the perfect lattice. These will be termed *point defects*. At the end of this lecture, we will give a short overview on extended defect structures, too.

Point defects may be seen as the simplest kind of lattice disorder. However, as we will see, this simple class of defects already includes a significant amount of different defect structures implying a wide diversity of physical effects. Fig. 2 illustrates different types of point defects for the example of a rock salt structure.

3.1 Types of lattice disorder

The most common kind of a point defect is an empty lattice site that is supposed to be occupied in the ideal crystal structure. Such a defect is called *vacancy*.

According to the conservation rules listed in section 2.2, a vacancy cannot be formed by removing an atom (or ion) from a lattice site without conserving mass. (In other words, the removed atom has to move somewhere.) One way to accommodate the removed atom is to move it to the boundary of the solid, where it can occupy a vacant lattice site. This type of disorder is called ***Schottky-disorder*** (Fig. 2b).

Schottky-defects are typically formed close surfaces or grain boundaries. Afterwards the resulting vacancies diffuse into the solid and distribute statistically. The involved diffusion process requires a sufficient mobility of vacancies and ions. Therefore, elevated temperatures are often necessary to allow equilibration of the lattice.

In order to conserve the lattice site ratio (and charge neutrality), Schottky-defects typically have to be generated in equal numbers for all sublattices. For the example of a rock salt structure, this means an equal number of Schottky-defects on anion and cation sites. Typical examples for materials showing Schottky-disorder are chromium oxide, Cr_2O_3, as well as the perovskite compound $SrTiO_3$ which will be discussed in detail in the remainder of this lecture.

In other materials, atoms (or ions) can also sit in between the regular lattice structure, occupying so-called *interstitial* lattice sites. This type of disorder is then called ***Frenkel-disorder*** (Fig. 2c). Frenkel-defects can be formed individually in each sublattice, as the resulting vacancy-interstitial pair is charge neutral and conserves mass.

Frenkel-defects, however, require sufficient *space* in the crystal lattice for the atoms to sit on interstitial lattice sites. Therefore, Frenkel-defects are unlikely in close-packed crystals, in particular, in close-packed oxides (with some exclusions), so that Schottky-disorder is often the dominant lattice disorder formalism.

On the other hand, materials showing Frenkel-disorder such as AgI or AgCl can show significant ionic conductivity (here Ag^+ conduction), as ions can move quickly through the solid via interstitial sites.

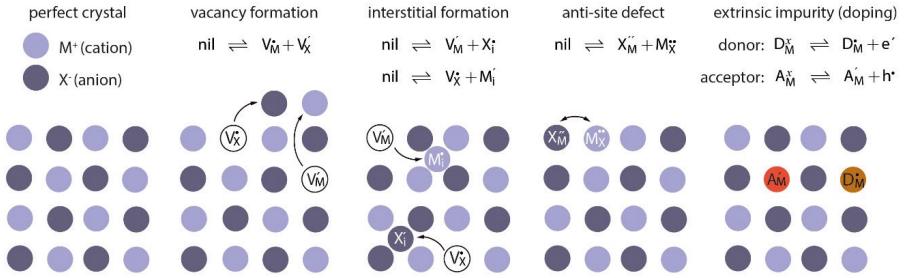


Fig. 2: *Various types of point defects in an ionic crystal for the example of a rock salt structure (a) such as AgCl. Point defects are generated by vacancy formation leaving behind unoccupied lattice sites. Atoms formerly sitting on that vacancy site can either move to the surface (b) or to an interstitial site (c). Moreover, point defects can include anti-site defects (d) and extrinsic impurities (e).*

Further types of point defects include so-called ***anti-site defects*** (Fig. 2d). In this case, two species of two different sublattices switch their positions. As a result, one receives one point defect in each sublattice. In ionic crystals, anti-site defects are often not favorable because of the involved electrostatic energies. In particular, anti-site defects between anion and cation sublattice have quite high formation enthalpies. In covalent crystals, however, where the on-site charges are small, e.g. in III-V semiconductors, anti-site defects can be an important type of disorder. In ionic binary and ternary compounds, however, the major types of intrinsic ionic disorder are mainly Frenkel-disorder and Schottky-disorder.

Another important family of point defects are ***extrinsic impurities***. This includes unintended contaminations (for instance during crystal growth) but also intended chemical doping or substitution. As known from semiconductor physics, extrinsic dopants can be used to vary the electronic properties of materials in a wide range. Dopants are typically included on actual lattice sites replacing the original atom (or ion) in the crystal lattice (Fig. 2e).

Generally, one distinguishes donor-type dopants bringing an extra electron into the solid, and acceptor-type dopants bringing one electron less than the original compound into the solid. In semiconductors, acceptor-type doping results in *p*-type conduction. However, as we will see, in complex oxides acceptor-type dopants are mostly compensated by ionic defects (oxygen vacancies).

3.2 Special character of oxides – oxygen exchange with ambient atmosphere

Among the ionic crystals, oxides have a particularly interesting property: the anion sites are occupied with oxygen ions. As we all know, oxygen is a major compound of the ambient atmosphere. Therefore, under some circumstances oxides have the ability to exchange oxygen with the ambient atmosphere.

This includes the incorporation of oxygen ions from the surrounding into the solid as well as the release of oxygen from the solid into the gas phase. These oxygen exchange reactions are of particular importance for many applications of complex oxides, such as for oxygen sensing, memristive devices and fuel cell applications.

The major consequence of oxygen exchange with the ambient atmosphere is that oxides incorporate oxygen vacancies as intrinsic ionic vacancies. If oxygen exchange with the ambient atmosphere is possible, the concentration of oxygen vacancies is not only dependent on temperature (cf. eq.(7)), but also on the ambient oxygen partial pressure. This results in an *intrinsic oxygen non-stoichiometry* in oxides, often denoted as δ. δ reflects the deviation in oxygen stoichiometry from its nominal value (e.g. $SrTiO_{3-\delta}$ or $(La,Sr)CoO_{3-\delta}$). In cobaltates, δ can be as large as 0.5.

In most oxides, oxygen vacancies are electronically active, in the sense that oxygen vacancies are charged. This is because oxygen ions are typically doubly ionized, O^{2-}. When they leave the solid in the course of an oxygen exchange reaction, they form gaseous charge neutral oxygen molecules, O_2. The two missing electrons per ion are left within the solid. Thus, oxygen vacancy formation commonly results in extra free electrons in the system. As a result, oxides with a high concentration of oxygen vacancies (reduced state) can exhibit high electrical conductivity, while the same compound with low amount of oxygen vacancies (oxidized state) can be a perfect insulator.

In addition, oxides tend to induce intrinsic lattice disorder in order to compensate extrinsic dopants and impurities. This is in particular the case for acceptor-type dopants. As a result, *p*-type conduction is rarely observed in oxides. In contrast, acceptor-type doping results in increased oxygen vacancy concentrations.

Therefore, rather the ionic conductivity than hole-type conduction can be tuned via acceptor-type chemical doping. This principle is for example exploited to tune and to optimize the ionic conductivity of oxygen ion conductors, such as ZrO_2 and CeO_2 used as electrolytes in solid oxide fuel cells.

In ionic crystals*, ionic charge compensation* mechanisms are as important as chemical doping. Electronic charge carrier densities are much more sensitive to atomic disorder (ionic defects) as compared to covalent solids. One and the same complex oxide compound can therefore transit from perfectly insulating behavior all the way to metallic behavior depending on its defect structure and in particular depending on its oxygen non-stoichiometry.

3.3 Defect notation

Defect formation processes often involve electrical charge, such as the generation and annihilation of electrons and electron holes or the valence change of ions. Therefore, it is useful to consider the relative charge of a defect or species with respect to the ideal crystal lattice. A vacancy on a cation (anion) lattice site represents a missing positive (or negative) charge for the crystal lattice. Relative to the ideal structure, a vacancy thus carries a negative (cation vacancy) or a positive relative charge (anion vacancy).

Typically, the convention introduced by Kröger and Vink (1953) [5] is used for defect notation and the handling of excess charges. The main features of Kröger-Vink notation are illustrated in Fig. 3 for the important example of an oxygen vacancy in an ionic oxide.

In Kröger-Vink notation, the (defect) species of interest is addressed by the main index, typically by the symbol of the chemical element. For the particular case of a vacancy, the notation 'V' is used.

A subscript refers to the addressed lattice site for a particular defect. Here, the chemical symbol of the original element sitting on the lattice site is used, e.g. `O` for an oxygen lattice site (Fig. ***3***). For the particular case of an interstitial site, the notation `i` is used.

A superscript indicates the net excess charge of the defect with respect to the original lattice charge, i.e. the charge deviation from the undisturbed, ideal crystal lattice. A bullet ($^{\bullet}$) accounts

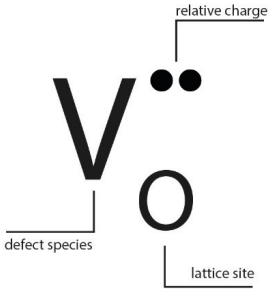


Fig. 3: *Kröger-Vink notation for the example of an doubly ionized oxygen vacancy. The main index reflects the addressed species (here V=vacancy), the subscript the addressed lattice site (O=oxygen), the superscript the relative charge of the defect ($^{\bullet\bullet}$=+2 relative charge).*

for one positive relative charge, a prime (') accounts for a negative relative charge. A neutral (or isovalent) defect is indicated by a cross (^{x}) or it is left without superscript.

For the example of an oxygen vacancy, the relative charge is doubly positive, because in the real lattice oxygen ions are typically doubly ionized (O^{2-}). An oxygen vacancy thus represents two missing negative elementary charges. Thus, an oxygen vacancy is denoted as $V_O^{\bullet\bullet}$ in Kröger-Vink notation.

Kröger-Vink notation is used also in chemical reaction equations. It is a common practice to omit *normal* components of the perfect crystal from the equilibrium reaction and to show only the defect species. The starting point for the reaction is then represented by the symbol 'nil', meaning 'no defects' or 'undisturbed lattice'.

Excess (free) electronic defects are denoted as e' for electrons and $h^{\bullet}$ for electron holes. In order to address defect concentrations, one often uses squared brackets, e.g. $\left[V_O^{\bullet\bullet}\right]$. Here, we will use the symbol c_{def} with appropriate subscript, e.g. $c_{V_O^{\bullet\bullet}}$, in order to address the concentration of a certain defect 'def'. The electron (hole) concentration will be denoted as *n (p)*.

Further examples for the Kröger-Vink notation are given for the example of an ionic rock salt structure in Fig. 2 and for the particular case of $SrTiO_3$ in Fig. 9.

4 Electronic disorder

4.1 Intrinsic electronic disorder

The treatment of intrinsic electronic disorder in solids in the framework of lattice disorder and thermodynamics, as introduced here, is naturally analogous to the classical derivation in semiconductor physics (see [6], [7]). Therefore, electronic disorder is a good starting point for the discussion of ionic disorder following in the next section.

Many complex oxides are wide-band-gap insulators, typically with a band gap a few electron volts (e.g. 3.2 eV for $SrTiO_3$ or 5.4 eV for $LaAlO_3$). In this case, intrinsic electronic disorder involves the thermal excitation of electrons from the valence band (VB) into the conduction band (CB), this is the formation of electron-hole pairs (Fig. *4*). In a pure, stoichiometric semiconducting or insulating compound, this is the only source of electronic carriers.

The chemical reaction for the formation of electron-hole pairs (band gap excitation) can be written as

$$\mathrm{nil} \rightleftharpoons \mathrm{e}' + \mathrm{h}^{\bullet} \tag{14}$$

where e' represents an electron in the conduction band, and $h^{\bullet}$ is a hole in the valence band. Here, we use Kröger-Vink notation as introduced in section 3.3.

In equilibrium, the corresponding law of mass action reads

$$n \cdot p = K_{\mathrm{I}}(T) = K_{\mathrm{I}}^{0} \cdot e^{-E_{\mathrm{g}}/k_{\mathrm{B}}T} . \tag{15}$$

The appropriate formation enthalpy here is the band gap, E_g, of the semiconductor or insulator. The factor K_I^0 can be expressed as the product of the effective density of states at the valence band, N_V, and conduction band edge, N_C

$$K_{\mathrm{I}}^{0} = N_{\mathrm{V}}(T) \cdot N_{\mathrm{C}}(T) . \tag{16}$$

In the intrinsic case, where eq. (14) is the only significant source of electrons and holes, n and p must be equal in order to preserve charge neutrality

$$n = p = \sqrt{K_1^0} \cdot e^{-E_g/2k_BT}. \tag{17}$$

The band gap may be considered to be inherently temperature dependent, E_g=E_g(T). The temperature dependence of E_g is typically expressed as

$$E_g = E_g^0 - \alpha T\,, \tag{18}$$

where E_g^0 is the band gap at zero temperature, and α is its linear temperature coefficient. Note, that eq. (18) is formally similar to eq. (1) describing Gibbs free energy, G=H-TS. Therefore, E_g is sometimes referred to as free energy, E_g^0 as enthalpy, and α is the entropy of the intrinsic electron ionization (band gap excitation).

4.2 Extrinsic electronic disorder

Similar to semiconductors, insulators such as most complex oxides can be doped by extrinsic impurities. In oxides, dopants are typically introduced on the cation sites, such as the Ti-sites in TiO_2, the Zr-sites in ZrO_2, or the Sr-sites and Ti-sites in $SrTiO_3$. Similar to semiconductors, oxides are typically doped by aliovalent compounds.

For example, $SrTiO_3$ is commonly doped with La on Sr-sites, or Nb on nominal Ti-sites. These *donor-dopants* bring an extra electron into the lattice. This extra electron can either be localized on the impurity site, or it can be ionized (Fig. 4). In that case, the electron is excited into the conduction band of the oxide where it contributes to electronic conduction.

Typically, the ionization energy of an extrinsic donor-dopant is the range of a few meV (shallow donors, e.g. in $SrTiO_3$) up to several hundreds of meV (deep donors, e.g. in ZnO).

The ionization process of extrinsic donor-type dopants is expressed by

$$D^x \rightleftharpoons D^\bullet + e'\,. \tag{19}$$

Complex oxides can also be doped with aliovalent compounds of lower valence. These dopants thus have acceptor character. For instance, $SrTiO_3$ can also be doped with Fe, Mn or Al. These elements are typically incorporated on nominal Ti-sites, mainly because of their ionic radii. Hence, while La is too large to be incorporated on Ti-sites, the smaller Fe-ion fits into the lattice

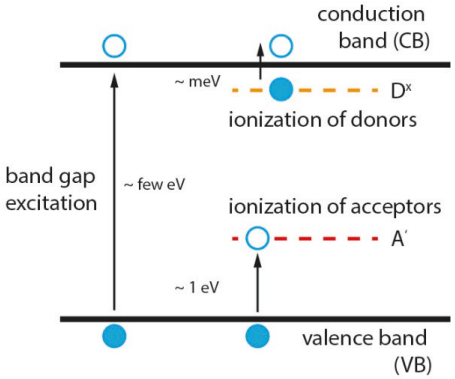


Fig. 4: *Electronic disorder in semiconductors and insulators. Electron-hole pairs are generated via band gap excitation processes. Moreover, donor-type impurities can ionize, adding an extra electron to the conduction band (CB), or acceptor-type impurities may ionize, accommodating one electron from the filled valence band (VB).*

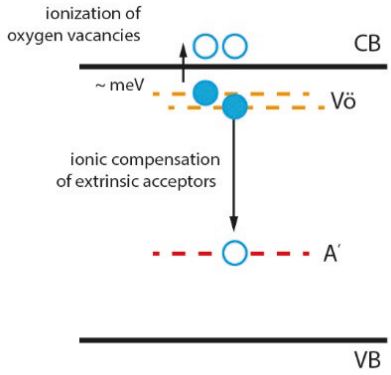


Fig. 5: *Ionic compensation of acceptor-type impurities by intrinsic oxygen vacancies. The electrons generated during oxygen vacancy formation may either be ionized into the CB, or they localize on acceptor-sites, thereby mainly suppressing hole formation in p-doped oxides.*

on a Ti-site. Therefore, materials with a similar valence can act as donors (such as La(3+) on Sr-sites) or as acceptors (such as Fe(3+) on Ti-sites).

Acceptor-type dopants thus contribute one electron less to the lattice than the original compound. As a result, an empty state in the valence band should be generated, such as known from semiconductor physics. The corresponding reaction equation reads

$$A^{x} \rightleftharpoons A' + h^{\bullet}. \tag{20}$$

However, as we will see later, acceptors in oxides are mostly deep acceptors and have ionization energies in the range of 1eV (Fig. 4). Therefore, at room temperature, acceptor-type impurities are often not ionized. Moreover, at elevated temperatures, oxides tend to include donor-type intrinsic defects, i.e. oxygen vacancies, into the lattice in order to compensate for acceptor-type impurities (Fig. 5). Therefore, *p*-type conduction is suppressed at room temperature in oxides. In most cases, hole contributions to the conductivity can only be observed at elevated temperatures.

5 Kinetic limitations

Ionic (or atomic) lattice disorder and the incorporation of point defects as discussed in section 3 requires the *migration of matter*, this is atoms or ions, through the solid. Therefore, ions have to be mobile in order to reach an equilibrium state.

Two important mechanisms control the mass transport in solids. These are *diffusion* and – for charged particles – *conduction*. While diffusion is driven by concentration gradients, conduction is driven by electric forces e.g. under applied bias.

As ions (and atoms) naturally are arranged on energetically favorable positions within the lattice, any motion will require overcoming energy barriers when leaving this favored position. Typically, this hopping process is thermally activated. Therefore, diffusion processes are accelerated at elevated temperature.

The mobility of ions and atoms is related to the diffusion coefficient

$$D = D_0 \exp\left(-\frac{Q}{k_B T}\right), \tag{21}$$

where the activation energy Q corresponds to the barrier height. The barrier heights and the diffusion coefficients can vary over orders of magnitude depending on the actual ions (or atoms) as well as on the solid under consideration. Strong electric fields may moreover lower energy barriers, so that electrical biasing may result in enhanced diffusion, too. Note that strong currents can also result in a local increase in temperature. This Joule heating is for instance important for the understanding of the memristive behavior of oxides.

As diffusion processes will be the focus topic of the following lecture by R. De Souza, we will not go into detail here. However, as an important aspect for the following discussion, we should note that the diffusion constants of different species (e.g. cations and anions) within the same solid could differ significantly. For example, oxygen anions typically migrate much faster in oxides than the associated cations, e.g. Sr^{2+} and Ti^{4+} in $SrTiO_3$. (The reason for this is commonly a vacancy diffusion mechanism as will be described in the next lecture.) Therefore, depending on the temperature range, one has to consider either anion migration, or cation migration or both in order to describe the established defect structure of a system at a given temperature.

Species with very low diffusion coefficients can be considered as immobile. Therefore, the sublattice of such a species is quasi-static and will not react on a change in temperature (or any other thermodynamic parameter). The defect concentrations are frozen in and will not equilibrate at a given temperature. The actual defect concentration established in the system then depends on the history of the sample, and the temperatures and atmospheres the solid has seen e.g. during crystal growth or previous annealing treatments.

6 An example: $SrTiO_3$

After a more general discussion, we will now turn to an important model material, namely $SrTiO_3$, which we will use as an example for determining and exploring lattice disorder effects in ionic oxides. After a short introduction into the basic material properties, the defect structure and its implications will be discussed for the cases of acceptor-doped and donor-doped $SrTiO_3$. A nice overview on this topic can be found also in Ref. [8]

6.1 General properties of $SrTiO_3$

$SrTiO_3$ is one of the most important materials in the family of perovskite oxides. $SrTiO_3$ is not only used as functional material in oxygen sensors, memristive devices, electro-chemical cells and transistors, but it also serves as a very important substrate material for epitaxial thin film growth of other perovskite compounds functionalized in novel oxide electronic devices and concepts.

$SrTiO_3$ is an ionic oxide with perovskite crystal structure, ABO_3, as displayed in Fig. 6. As an oxide, $SrTiO_3$ contains oxygen (O^{2-}) as anions[1], located in the center of each face of the unit cell (red). The corners of the unit cell, the so-called A-sites, are typically occupied by the larger cation of the compound. In the case of $SrTiO_3$, this is Sr^{2+} (light blue). The center of the unit cell, the so-called B-site, is then occupied by a smaller second cation. In the case of $SrTiO_3$,

[1]While this is the case in most perovskite crystals, perovskites also exists as non-oxides, which are used as high-efficiency absorber materials in solar cells. These compounds usually contain Cl^-, I^-, or Br^- as anions and more complex cationic groups, e.g. $(CH_3NH_3)PbCl_3$.

this is Ti^{4+} (dark blue). The B-site cation is surrounded by an oxygen ion octahedral formed by the six oxygen ions (red shaded area in Fig. 6).

While the ionic picture of $SrTiO_3$ is sufficient to understand many material properties, it is indispensable to approach the material also in terms of band structure. From this view, stoichiometric $SrTiO_3$ is a d^0 band insulator with a band gap of 3.2 eV at 0 K. Because of the large band gap, $SrTiO_3$ single crystals are typically transparent. However, in the non-stoichiometric case, crystals can appear in dark blue and even blackish. As we will see, this is a result of free carriers induced via extrinsic dopants and/or oxygen vacancies.

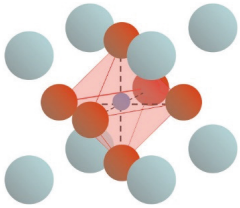


Fig. 6: *Perovskite unit cell: In the case of $SrTiO_3$, the A-sites (corners of the unit cell) are occupied by Sr^{2+} ions (light blue), while the B-site (center of the unit cell) is occupied by a Ti^{4+} ion (dark blue). The B-site is surrounded by an oxygen octahedral (one oxygen anion (O^{2-}, red) on each face of the unit cell).*

Similar to many other transition metal oxides such as TiO_2, $BaTiO_3$, etc., the valence band of $SrTiO_3$ corresponds mainly to O 2p states, while the conduction band originates mainly from Ti 3d states. Consistent to the ionic picture, the filled oxygen 2p states in the VB correspond to the formation of O^{2-} ions, while the empty Ti 3d states in the CB indicate a nominal Ti valence state of 4+.

$SrTiO_3$ has a high dielectric constant (ε_r=300 at 300K up to a few thousand at low temperature) and has been tested for use as dielectric layer in transistors. Sometimes, $SrTiO_3$ is even referred to as incipient ferroelectric material. Ferroelectricity can be stabilized e.g. by epitaxial and mechanical strain, however, without practical use. For ferroelectric applications rather the related perovskite compound $BaTiO_3$ is used, which shows much stronger and more stable polarization.

The large dielectric constant implies a large electrostatic screening length. This effect is utilized in the design of novel 2D superconductor in doped-superlattice structures [9]. Moreover, this results in significant effects of space charge layers at surfaces and interfaces of $SrTiO_3$ as will be discussed in section 7.

6.2 Defect concentrations in thermodynamic equilibrium

In terms of lattice disorder, $SrTiO_3$ shows exclusively Schottky-type of disorder, while Frenkel-defects, i.e. interstitials, are absent. This can be understood from Fig. 7, again showing a single unit cell of $SrTiO_3$. Here, the real ionic radii of the compounds are considered. One oxygen ion has been removed from the front face of the unit cell in order to allow a clear view on the Ti ion. Oxygen and strontium ions have a rather similar size, while the Ti ion in the center of the

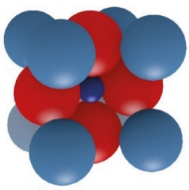


Fig. 7: *$SrTiO_3$ unit cell considering the real ionic radii of the ions. Sr^{2+} and O^{2-} are rather similar in size, while $Ti^{4+\,(3+)}$ is much smaller. The structure is close-packed inhibiting the formation of Frenkel-defects.*

oxygen octahedral is much smaller. The perovskite structure is close-packed, so that the particular ions *touch* each other. Hence, there is no space for interstitial sites in the crystal lattice and we can restrict our discussion of defect structure in $SrTiO_3$ to Schottky-disorder.

Moreover, we can exclude anti-site defects because of the very different size, charge and coordination of the various compounds.

The various types of Schottky-disorder that remain to be considered in $SrTiO_3$ are displayed in Fig. 8. These are oxygen vacancies (a), extrinsic doping (b), strontium vacancies (c) and titanium vacancies (d). The corresponding defect notations relevant to the following discussion are listed in Kröger-Vink notation in Fig. 9.

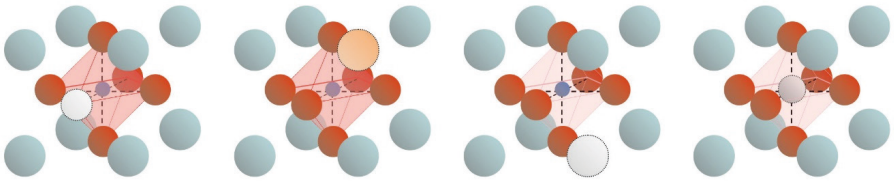


Fig. 8: *Schottky-defects in $SrTiO_3$: (a) vacant oxygen site (oxygen vacancy), (b) substitution of Sr^{2+} by e.g. La^{3+} (extrinsic impurity/doping), (c) vacant Sr-site (strontium vacancy), (d) vacant Ti-site (Ti-vacancy).*

At room temperature, all ions are immobile in $SrTiO_3$. Therefore, the concentrations of all ionic vacancies are frozen, unless strong local electric fields are applied or the material is heated up locally to much higher temperatures. Hence, at low temperature only electronic equilibria are active, resulting in merely electronic disorder as discussed in sec. 4.

The formation and equilibration of ionic disorder requires significantly fast diffusion of the vacancies. The diffusion coefficients of $V_O^{\bullet\bullet}$ and $V_{Sr}^{''}$ in $SrTiO_3$ have been studied extensively by various experimental techniques and by simulation. It turned out that oxygen vacancies show a rather high mobility in the perovskite lattice and a low activation energy of diffusion (0.6 to 1.0 eV). In contrast, strontium vacancies exhibit a very low mobility and a high activation energy of diffusion (2.5 to 3.5 eV). For Ti-sites, the barrier heights are even larger as we will discuss below.

chemical species	Kröger-Vink Notation
incorporated O^{2-} anion	O_O^x
incorporated Sr^{2+} cation	Sr_{Sr}^x
incorporated Ti^{4+} cation	Ti_{Ti}^x
oxygen vacancy	$V_O^{\bullet\bullet}$
strontium vacancy	V_{Sr}''
titanium vacancy	V_{Ti}''''
free electron	e'
free hole	$h^\bullet$
trivalent acceptor on Ti-site	A'_{Ti}
trivalent donor on Sr-site	$D^\bullet_{Sr}$

Fig. 9 : Kröger-Vink notation of the relevant defect species in $SrTiO_3$.

Depending on the considered temperature range (and mobility of the ions), one thus has to consider a different number of active crystal sublattices that contribute to lattice disorder in thermodynamic equilibrium.

6.2.1 Equilibration of the oxygen sublattice

Above approximately 500°C, oxygen ions (and vacancies) get mobile. Moreover, the surface exchange of oxygen is accelerated. For the incorporation of oxygen into the bulk, this includes

1) splitting of gaseous O_2 into surface-adsorbed O* atoms
2) ionization of these adsorbed oxygen atoms (O^{*2-} formation)
3) incorporation of adsorbed ions into the bulk

or the opposite reactions for the release of oxygen from the bulk [10]. In other words, the oxygen sublattice $SrTiO_3$ starts to equilibrate with the surrounding atmosphere, characterized by the oxygen partial pressure, pO_2.

The removal of oxygen from the lattice is called *reduction* and can be written in Kröger-Vink notation as

$$O_O^x \rightleftharpoons V_O^{\bullet\bullet} + 2e' + \tfrac{1}{2} O_2(g) . \tag{22}$$

An oxygen ion leaves the crystal into the gas (g) phase and forms a neutral molecule. For this, it releases two electrons to the lattice and leaves behind a double-positively charged oxygen vacancy, $V_O^{\bullet\bullet}$.

The opposite reaction of oxygen incorporation can be written as

$$\tfrac{1}{2} O_2(g) + V_O^{\bullet\bullet} \quad \rightleftharpoons \quad O_O^x + 2h^{\bullet} . \tag{23}$$

Here, (half) a gaseous O_2 molecule is incorporated on a vacant oxygen lattice site, resulting in an now-occupied regular oxygen lattice site, O_O^x. For the required ionization of the oxygen ion, this reaction consumes two electrons that have to be provided by the lattice, here indicated by the formation of two holes in the valence band, $2h^{\bullet}$.

Hence, if oxygen exchange with the surrounding atmosphere is possible, both reactions will take place and will strive for thermodynamic equilibrium. Therefore, we can use the law of mass action of both reactions, in order to determine the equilibrium concentrations of oxygen vacancies, electrons, and holes.

The law of mass action reads

$$c_{V_O^{\bullet\bullet}} \cdot n^2 \cdot (pO_2)^{1/2} = K_{red}^0 \cdot e^{-(\Delta H_{red}/k_B T)} \tag{24}$$

for the reduction reaction and

$$\frac{p^2}{c_{V_O^{\bullet\bullet}} \cdot \left(pO_2\right)^{1/2}} = K_{ox}^0 \cdot e^{-(\Delta H_{ox}/k_B T)} . \tag{25}$$

Here, ΔH_{red} and ΔH_{ox} denote the reaction enthalpies for the reduction and oxidation of $SrTiO_3$. Later, we will discuss how these values can be determined experimentally. Here, we replaced the activity or concentration of gaseous molecules by the oxygen partial pressure, pO_2.

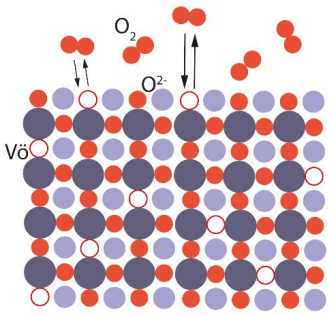


Fig. 10: *Oxygen exchange between oxide and ambient atmosphere. Oxygen can be incorporated from the gas phase into the solid (and vice versa) if the surface reaction kinetics as well as the diffusion kinetics are sufficiently fast. Typically, this is achieved at elevated temperatures (~500°C).*

Multiplication of eqs. (24) and (25) yields

$$n \cdot p = \sqrt{K^0_{\mathrm{red}} K^0_{\mathrm{ox}}}\, \exp\left(-\frac{\Delta H_{\mathrm{ox}} + \Delta H_{\mathrm{red}}}{2 k_{\mathrm{B}} T}\right), \tag{26}$$

which is identical to the result obtained for electron-hole pairs (eq.(15)). We can therefore identify

$$E_{\mathrm{g}} = \frac{\Delta H_{\mathrm{ox}} + \Delta H_{\mathrm{red}}}{2}, \tag{27}$$

hence yielding a relation between ΔH_{red}, ΔH_{ox} and the band gap, E_{g}, of $SrTiO_3$.

6.2.2 Equilibration of the cation sublattice

If we further increase the temperature of the system – above approximately 1000°C – also the cation lattice starts to equilibrate, as the mobility of the cations is now sufficiently high. Note, however, that lattice equilibration of the cation sublattice is still very sluggish on the macroscopic scale at these temperatures. Thus, it takes a significant amount of time, this is a few hours up to days to equilibrate the cation sublattice, while the oxygen sublattice equilibrates within micro- to milliseconds (!).

In the course of equilibration, the cation sublattice incorporates vacancies. As we now consider cations (instead of anions in the oxygen case), a cation vacancy represents a missing positive charge in the real lattice. Thus, a cation vacancy carries a negative relative charge with respect to the lattice, a Ti-vacancy, v'''_{Ti}, then is four times charged, while a Sr-vacancy is doubly charged, v''_{Sr}.

Following the conservation rules introduced in section 2.2, cation vacancies can be induced by removing an entire $SrTiO_3$ unit cell from the lattice, forming one v'''_{Ti}, one v'''_{Ti} and three $v_{\ddot{\mathrm{O}}}$ at a time

$$\mathrm{Sr}^{\mathrm{x}}_{\mathrm{Sr}} + \mathrm{Ti}^{\mathrm{x}}_{\mathrm{Ti}} + 3\mathrm{O}^{\mathrm{x}}_{\mathrm{O}} \rightleftharpoons \mathrm{V}''_{\mathrm{Sr}} + \mathrm{V}'''_{\mathrm{Ti}} + 3\mathrm{V}^{\cdot\cdot}_{\mathrm{O}} + (\mathrm{SrTiO_3}). \tag{28}$$

This is the so-called *Schottky-equilibrium*.

In the case of $SrTiO_3$, it turns out that only oxygen vacancies and strontium vacancies are formed. Titanium vacancies are less likely, as their formation is energetically more costly. As one can see qualitatively in Fig. 7, this is because the highly charged Ti^{4+}-cations are squeezed into the center of the oxygen anion octahedral, where they are tightly surrounded by the oxygen

anions. The resulting strong Coulomb interactions result in a much higher defect formation energy for Ti-vacancies than for strontium vacancies. Moreover, the energy barriers for Ti-diffusion are rather high because of this coordination. Thus, also Ti-diffusion is suppressed, inhibiting the equilibration of the Ti-sublattice even at very high temperatures.[2]

Therefore, rather a *partial Schottky-equilibrium* has to be considered for $SrTiO_3$ which only addresses the Sr-cation sublattice

$$Sr_{Sr}^{x} + O_{O}^{x} \rightleftharpoons V_{Sr}'' + V_{O}^{\cdot\cdot} + SrO_{s.p.}\,. \tag{29}$$

Here, $SrO_{s.p.}$ represents a strontium oxide secondary phase which is excorporated on external or internal surfaces (i.e. grain boundaries) of the crystal. The formation of secondary phases is necessary in order to conserve mass, charge and the lattice site ratio, which would be changed if strontium moved onto a regular lattice site on the surface of the crystal without being accompanied by a Ti-ion. The secondary phase formation often observed on the surface of $SrTiO_3$ in the form of precipitates. The law of mass action of the *partial Schottky-equilibrium* reads

$$c_{V_{Sr}''} \cdot c_{V_{O}^{\cdot\cdot}} = K_S^0 \exp\left(-\frac{\Delta H_S}{k_B T}\right). \tag{30}$$

6.2.3 Electro neutrality

Before finally determining the equilibrium lattice disorder in $SrTiO_3$, we need to introduce another boundary condition, this is charge neutrality. While each defect formation reaction itself requires to be charge neutral (conservation rule 2), also the solid itself has to be charge neutral. In particular, as multiple defect formation reactions are involved, their coupling has to be mediated via the local charge neutrality condition – which is valid locally in the absence of electric fields.

As discussed before, ionic defects carry charge and thus contribute to the charge balance of the solid. The electron neutrality condition thus reads

$$n + 2c_{V_{Sr}''} + 4c_{V_{Ti}''''} + c_{A'} = p + 2c_{V_{O}^{\cdot\cdot}} + c_{D^{\bullet}}\,. \tag{31}$$

Here, we added all negatively charged defects, including cation vacancies and ionized acceptor-type dopants (A'), on the left hand side. On the right hand side, we added all positively charged defects including oxygen vacancies and ionized donor-type dopants ($D^{\bullet}$). As stated above, the contribution of Ti-vacancies can be neglected in most cases.

6.3 Acceptor-doped $SrTiO_3$

We now consider acceptor-doped $SrTiO_3$ (i.e. $c_{D^{\bullet}} = 0$). Acceptor-doping can be achieved intentionally, e.g. by doping with Fe, Al, or Mn. However, even *nominally undoped* $SrTiO_3$ is always slightly acceptor-doped. This is because the most common contaminations naturally incorporated during single crystal growth are almost exclusively acceptor-type. Typically, undoped $SrTiO_3$ single crystals have an impurity level of 10-100 ppm, corresponding to $c_{A'} \approx 1\times10^{17} - 1\times10^{18}$ cm^{-3}.

[2] Interestingly, this is different in $BaTiO_3$. The large Ba^{2+}-ions expand the lattice, and thus the oxygen octahedral. Therefore, the ionic bonds of Ti-ions are weaker than in $SrTiO_3$. As a result, Ti-vacancies are the preferred cationic defect in $BaTiO_3$. [3]

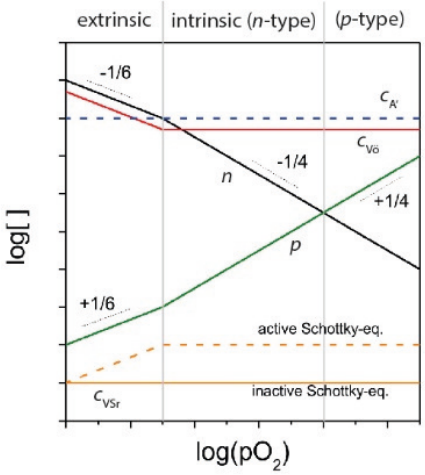


Fig. 11: *The Brouwer diagram of acceptor-doped $SrTiO_3$ illustrates the defect concentrations as a function of ambient pO_2 on double-log-scales. Taken from [11].*

We will now discuss how the lattice of acceptor-doped $SrTiO_3$ reacts on a change in oxygen partial pressure in the surrounding atmosphere. For this, it is required that oxygen exchange with the ambient atmosphere is allowed, i.e. temperatures above 500°C are considered (see sec. 6.2). At this temperature, all acceptor-dopants can be considered to be ionized.

Intuitively, one may expect that as a result of oxygen exchange with the ambient atmosphere, the oxygen vacancy concentration incorporated in the lattice varies as a function of pO_2. As we will see, this is to some extend indeed the case. Moreover, due to the coupling of oxygen vacancies and electronic charge carriers via eqs. (24) and (25), also the electronic charge carrier concentrations and thus the conductivity of the oxide vary with pO_2.

Fig. 9 shows the defect concentrations established in this system at a given temperature as a function of ambient atmosphere on double logarithmic scales. Such figures are often called *Brouwer diagrams*.

The characteristic result obtained for $SrTiO_3$ is representative for many acceptor-doped oxides, such as TiO_2, $BaTiO_3$, etc. We will derive this behavior based on the previous considerations. Starting at very low oxygen partial pressure, we will increase the pO_2 stepwise.

At low oxygen partial pressure, the concentration of electrons and oxygen vacancies has to be large according to the reduction reaction eq. (24). Thus, if we decrease the oxygen partial pressure far enough, it is reasonable to assume that oxygen vacancies are the major positively charged defect in the system, and electrons are the major negatively charged defect. Thus, the charge neutrality condition reads

$$n = 2c_{V_O^{\cdot\cdot}}, \tag{32}$$

with $n \gg 2c_{V_{Sr}''} + c_{A'}$ and $c_{V_O^{\cdot\cdot}} \gg p$. This is the *extrinsic regime*, as the carrier density is directly determined by the amount of oxygen non-stoichiometry.

Plugging in eq. (32) in eq. (24), we obtain

$$c_{V_O^{\cdot\cdot}} \propto pO_2^{-1/6} \qquad \left[\text{i.e. } \log\left(c_{V_O^{\cdot\cdot}}\right) \propto -\frac{1}{6}\log\left(pO_2\right)\right]. \tag{33}$$

Directly it follows

$$n \propto pO_2^{-1/6} \tag{34}$$

and using the oxidation reaction or band gap excitation

$$p \propto pO_2^{+1/6}. \tag{35}$$

Note, that since n is large, p is small in order to keep the product $(n \cdot p)$ constant.

With increasing pO_2, the concentration of oxygen vacancies will thus decrease (see Fig. 11). At a certain pO_2, n will be comparable to or even drop below the acceptor-level. Thus, our assumption of electrons being the dominant negatively charged defect species is no longer valid. Instead, the charge neutrality condition now reads

$$c_{A'} = 2c_{V_O^{\cdot\cdot}}. \tag{36}$$

The acceptor concentration, however, is a constant and does not change with pO_2. Therefore, also the oxygen vacancy concentration is virtually pinned by the acceptor-dopants. Further oxidation, i.e. increase of the pO_2, will not result in a further decrease of the oxygen vacancy concentration. As a result, one always maintains a minimum concentration of oxygen vacancies in acceptor-doped $SrTiO_3$. Even more, the concentration of oxygen vacancies can be tuned via the acceptor- dopant concentration (eq. (36)). The same principle is used in e.g. $Y:ZrO_2$ to optimize ion conduction in solid oxide electrolytes.

Again, we plug in the charge neutrality condition into the mass action law of the oxygen exchange reactions (same order as above) and obtain

$$c_{V_O^{\cdot\cdot}} \propto pO_2^{0}, \; n \propto pO_2^{-1/4} \text{ and } p \propto pO_2^{+1/4}. \tag{37}$$

Hence, while the oxygen vacancy concentration is constant, the electron concentration decreases further, now with a changed slope of -1/4 in the Brouwer diagram. This is the so-called *intrinsic regime* as n and p are mainly determined by electronic defect equilibria in this regime.

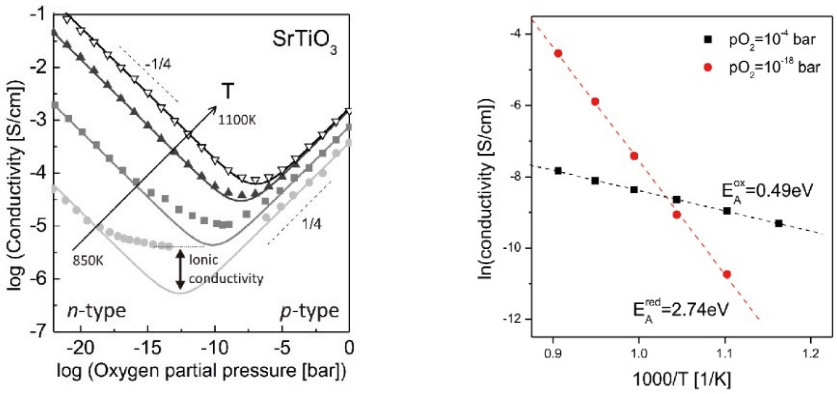


Fig. 12: *High temperature equilibrium conductance (HTEC) of $SrTiO_3$ revealing the typical slopes of +/-1/4 in the intrinsic conductance regime of complex oxide compounds (a). The temperature dependence in the p-type (10^{-4} bar) and n-type (10^{-18} bar) regime shows a typical Arrhenius-type behavior (b). The slopes can be correlated to the reaction enthalpies ($\Delta H_{red,ox} = 2E_A^{red,ox}$). Adopted from Refs. [11, 12].*

As the hole concentration further increases within the intrinsic regime, one finds a cross-over from dominant *n*-type conduction to dominant *p*-type conduction in oxidizing conditions.

Summarizing the discussion above, the nominal *insulator* $SrTiO_3$ can be an *n*-type *conductor* when lattice disorder has been established in reducing atmosphere. At elevated temperature, $SrTiO_3$ can even be a *p*-type *conductor* (!) in oxidizing conditions. Note, however, that *p*-type conduction is suppressed at lower temperatures such as room temperature, as electron holes freeze out significantly upon cooling, because of the large ionization energy of the acceptors. Holes are thus trapped on the acceptor-sites, where they localize and deionize the acceptor.

At elevated temperature, however, both contributions can be measured experimentally. Fig. 12 (a) shows experimental data obtained for an undoped (i.e. impurity acceptor-doped) $SrTiO_3$ single crystal. The characteristic slopes of the intrinsic regime, i.e. -1/4 for the *n*-type conduction regime and +1/4 for the *p*-type regime are clearly observed.

Close to the electronic conduction minimum, an additional conduction contribution can be observed owing to ionic conduction. From the activation energy of conduction in the intrinsic *n*-type (*p*-type) region, the reaction enthalpy for reduction (oxidation) can be determined as shown in Fig. 12 (b). We can identify

$$E_g = \frac{\Delta H_{ox} + \Delta H_{red}}{2} = \frac{2E_A^{ox} + 2E_A^{red}}{2} \approx 3.2\text{eV}\,, \tag{38}$$

as expected from eq. (27).

The oxygen vacancy concentration in acceptor-doped $SrTiO_3$ is always significantly high. As a result, in equilibrium the concentration of strontium vacancies is always very low, according to eq. (30). Therefore, the Brouwer diagram does not change a lot (cf. dashed line in Fig. 11) when the temperature is further increased and cations have to be considered mobile. This will change dramatically in the case of donor-doped $SrTiO_3$, which will be discussed next.

6.4 Donor-doped STO

We now consider donor-doped $SrTiO_3$ (i.e. $c_{A'} = 0$) in a similar way as done before for the acceptor-type case. Donor-doping is typically achieved by the admixture of lanthanum or niobium. Both, La:$SrTiO_3$ and Nb:$SrTiO_3$ single crystals are commercially available with dopant levels ranging from a few ppm up to a few atomic percent.

At first, we will consider that only the oxygen sublattice contributes to lattice disorder (i.e. temperatures of about 500°C). Afterwards, it will be discussed what consequences appear if also the cations, this is the strontium sublattice, equilibrates.

Fig. 13 shows Brouwer diagrams for donor-doped $SrTiO_3$ for the case of inactive cation sublattice (a) and for the case of an active cation sublattice (b).

Starting again at very low pO_2, we can again assume that electrons and oxygen vacancies are the major defect species in the system. Again, we thus start with the *extrinsic regime*, identical to the acceptor-doped case. The pO_2-dependence of electron density and oxygen vacancy concentration again follows a -1/6-power law.

At a certain pO_2 value, the oxygen vacancy concentration will then be comparable to or drop below the donor level, $c_{D'}$. Hence, oxygen vacancies are no longer the dominant positively charged defect. Instead of eq.(32), the charge neutrality condition now reads

$$n = c_{D'} = \text{const.} \tag{39}$$

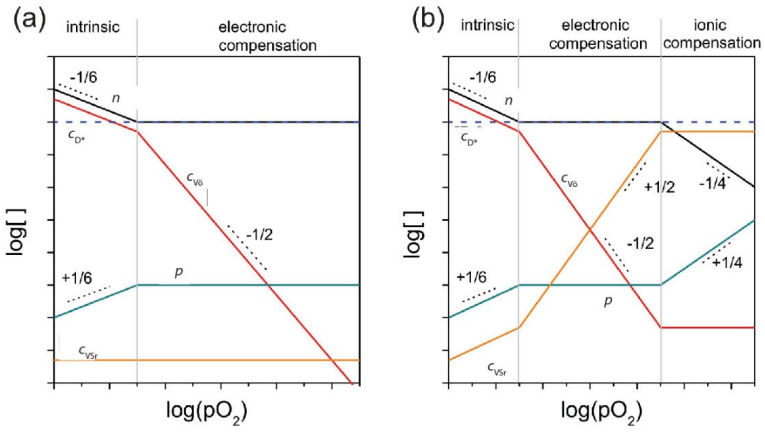


Fig. 13: *Brouwer diagram for donor-doped $SrTiO_3$: (a) inactive Schottky-equilibrium; (b) active Schottky-equilibrium. After [11, 13].*

Thus, in the case of donor-doped $SrTiO_3$, the electron concentration is pinned by the extrinsic donor, resulting in a pO_2-independent *plateau-region* in the electron concentration. The donor-states are thus compensated by electrons, which is called *electronic compensation.* This corresponds to the standard case known from classical semiconductors. Therefore, donor-doped $SrTiO_3$ is a good *n*-type conductor, often used e.g. as bottom electrode in epitaxial perovskite thin film structures.

As *n*=const., the law of mass action of the reduction reaction yields

$$c_{V_O^{\cdot\cdot}} \propto pO_2^{-1/2} \quad . \tag{40}$$

Hence, while the electron concentration is constant, the oxygen vacancy concentration drops dramatically when further increasing the pO_2. $c_{V_O^{\cdot\cdot}}$ reaches values that are orders of magnitude smaller than in acceptor-doped $SrTiO_3$. As a result, oxygen ion diffusion is dramatically suppressed in the donor-doped case, while it is much faster in the acceptor-doped case.

Unlike in the acceptor-doped case, the Brouwer diagram of donor-doped $SrTiO_3$ changes significantly if the strontium sublattice contributes to lattice disorder. Considering eqs. (29) and (30), a *decreasing* oxygen vacancy concentration is accompanied by an *increasing* concentration of strontium vacancies. One thus gets

$$c_{V_{Sr}''} \propto pO_2^{+1/6} \tag{41}$$

in the extrinsic regime, and

$$c_{V_{Sr}''} \propto pO_2^{+1/2} \tag{42}$$

in the plateau-region. For an active Schottky-equilibrium, the concentration of strontium vacancies is hence drastically enhanced as compared to the acceptor-doped case.

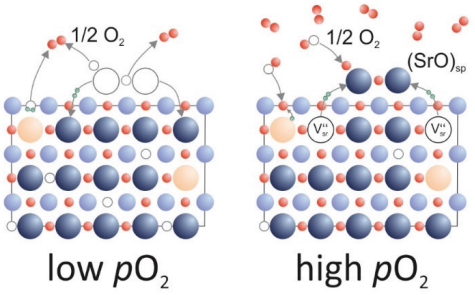


Fig. 14: *The partial Schottky equilibrium in (donor-doped) $SrTiO_3$ balances ionic and electronic compensation of extrinsic donors: While electronic compensation dominates at low pO_2, strontium vacancy formation in high oxygen pressures triggers ionic charge compensation.*

At a certain pO_2, the strontium vacancy concentration will cross and exceed the constant electron concentration. Hence, eq. (39) is no longer valid. Instead, the positive charge of the donors is then compensated by the negative charge of incorporated strontium vacancies

$$2c_{V_{Sr}''} = c_{D^{\cdot}} \,. \tag{43}$$

The amount of incorporated cation vacancies is thus comparable to the extrinsic donor concentration (i.e. up to a few atomic percent, depending on the doping level). This is the so-called *ionic compensation.*

Once ionic compensation is achieved, the strontium vacancy concentration settles and stays constant upon further increasing the ambient pO_2. In contrast, n is now decreasing following a -1/4-power law. Thus, upon further oxidation, the n-type conduction is diminished. Donor-doped $SrTiO_3$ thus turns into an insulator upon oxidation at high temperature.

The process of strontium vacancy formation requires the formation of SrO secondary phases at the surface of the crystal as illustrated in Fig. 14. As a result, one often observes precipitate formation.

Via lattice disorder effects on can tune the electronic properties of donor-doped $SrTiO_3$ from metal-like conduction to insulating behavior. This effect is particularly important at surfaces and interfaces. Surface oxidation is for instance believed to be the major mechanism in oxygen sensors utilizing La:$SrTiO_3$ ceramics.

Similar to the donor-doped case, an ionic charge compensation effect has also been observed in the 2-dimensional electron gas formed at the interface of $SrTiO_3$ and $LaAlO_3$ [13,14], which has attracted a lot of attention for fundamental research in recent years. Moreover, strontium segregation effects are considered in the aging of complex oxide cathodes used in solid oxide fuel cells. [15]

As has been discussed, lattice disorder in the bulk of complex oxides can vary dramatically depending on the ambient atmosphere and temperature, but also on history of the sample, which may determine defect concentrations of species that are frozen in at lower temperatures. Moreover, it has been shown that different extrinsic doping of the same materials can lead to significantly different lattice disorder configurations.

By understanding and controlling lattice disorder in complex oxide materials one has the possibility to tune the material properties over a wide range, e.g. electronic properties can change from metallic conduction to insulating behavior by changing tiny amounts of lattice disorder, making this topic significantly important for many applications.

While we have so far concentrated on bulk properties, the following chapter will discuss in detail the defect structure near surfaces and interfaces, which may differ significantly from the bulk.

7 Grain boundaries and surfaces

In many applications, surfaces and interfaces play an important role. This is particularly the case for *thin films* as well as for *ceramics* having a large amount of grain boundaries. Therefore, it is an important question how lattice disorder is affected by the presence of surfaces, interfaces and grain boundaries.

In semiconductor materials, one typically observes band bending at the surface as a result of surface charges generating an electrostatic potential in the near-surface region. As it turns out, such surface charges may exist in ionic oxides, too.

In ionic crystals, however, surface charges and electrostatic potentials do not only affect the electronic band structure, but also the *ionic* structure – there will be "band bending" also for ionic defect species, in the sense that the electro-chemical potential for a particular defect will bend close to the surface. As a result, one observes inhomogeneous defect concentration profiles near ionic crystal surfaces, in particular accumulation and depletion of ionic defects. In order to discuss the space charge formation in oxides, we will again use the model material system of $SrTiO_3$ introduced in the forgoing section.

Already in the 90s in was realized that acceptor-doped $SrTiO_3$ ceramics exhibit an unusual capacitive contribution in addition to the expected bulk capacitance as characterized by impedance spectroscopy (Fig. 15). This additional capacitance was attributed to a grain boundary contribution. In particular, it was realized that the grain boundaries of acceptor-doped $SrTiO_3$ ceramics are positively charged, generating a space charge layer in which negatively charged defects accumulate, while positively charged defect are depleted (Fig. 15, right).

Surface charges are often termed *core charges* referring to the grain boundary core.

The width of the space charge layer (SCL), d_{SCL}, was estimated from the grain boundary capacitance, C_{SCL}, assuming a simple plate capacitor model

$$C_{SCL} = \varepsilon_0 \varepsilon_r \frac{A}{d_{SCL}} \Rightarrow d_{SCL} = \varepsilon_0 \varepsilon_r \frac{A}{C_{SCL}}, \tag{44}$$

where ε_r is the dielectric constant of $SrTiO_3$, and A the contact area of the electrodes used for the impedance experiment.

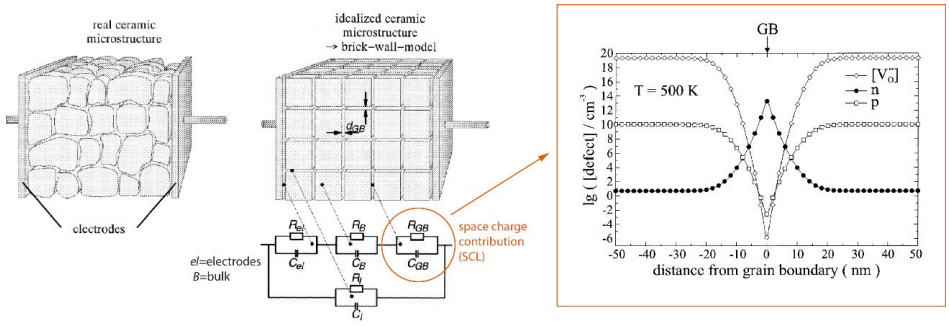


Fig. 15: *Space charge formation at grain boundaries in acc.-doped $SrTiO_3$-ceramics. Left: The additional capacitance of the space charge layer (C_{SCL}=C_{GB}) can be measured by impedance spectroscopy. Right: A positive surface charge results in depletion of oxygen vacancies and holes at the grain boundary, while electrons accumulate. Figs. taken from Ref. [16] (modified).*

Typically, values of a few tens of nanometers up to a few hundreds of nanometers were found for d_{SCL}, which can be correlated to the surface potential, Φ^0_{SCL}, established at the grain boundary core

$$d_{SCL} = \sqrt{\frac{\varepsilon_0 \varepsilon_r \Phi^0_{SCL}}{e c_{A'}}} . \tag{45}$$

Typically, Φ^0_{SCL} is in the range of 0.5-0.8 V.

At that time, the nature of the surface charge mediating space charge formation was not yet realized. Nevertheless, a proper description of the ionic constitution within the space charge layer was established (Fig. 15, right) [16, 17]. Only in recent years, the nature of the positive core charge at surfaces and grain boundaries was identified as excess oxygen vacancies. It turns out that the formation energy for oxygen vacancies at the surface of $SrTiO_3$ is about 1.4 eV lower than in the bulk (Fig. 16). Intuitively, this can be understood by the fact that in order to remove an oxygen ion from the surface less ionic bonds have to be broken than in the bulk.

As a result, oxygen vacancies favor being located right at the surface or at grain boundaries of $SrTiO_3$. This triggers a diffusion process for oxygen vacancies towards the surface, which is accompanied by space charge formation. Experimentally, this has been confirmed by O^{18} exchange diffusion experiments [18] (see following lecture) as well as by the measurement of the high temperature equilibrium conductance in epitaxial $SrTiO_3$ thin films [19].

For a general treatment of surface space charge layers, an implicit electrostatic problem has to be solved. For the sake of simplicity, often translation symmetry in two dimensions is assumed, reducing the problem to a 1-dimensional one.

Let x denote the distance from the surface of a $SrTiO_3$ single crystal. Any surface charge, Q_c (c=core), generates an electric field at the surface (x=0) as described by Gauss' law (Fig. 17)

$$E\big|_{x=0} = -\frac{d\phi}{dx}\Big|_{x=0} = \frac{Q_c}{\varepsilon_0 \varepsilon_r} . \tag{46}$$

The natural reaction of the material to an applied field is the redistribution of all mobile charge carriers in order to screen the field. In ionic crystals, this may include electrons, electron holes, but also oxygen vacancies and sometimes even cation vacancies (e.g. in $La:BaTiO_3$ ceramics).

Associated with this process, an electrostatic potential, $\phi(x)$, is established. Because of the electrostatic energy, $z_{def} e\phi(x)$, the mobile defect concentrations within the space charge region align with the potential. Here, z_{def} denotes the charge number of the defect. The concentrations within the space charge layer can (in approximation) be derived as

$$c_{def}(x) = c_{def}^{bulk} \exp\left(-\frac{z_{def} e\phi(x)}{k_B T} \right) . \tag{47}$$

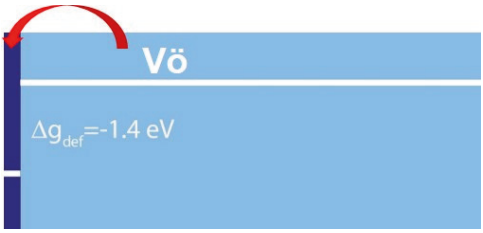


Fig. 16 : *Driving force for space charge formation at the surface (grain boundaries) of acceptor-doped $SrTiO_3$. The enthalpy for oxygen vacancy formation is about 1.4eV lower at the surface, as compared to the bulk, triggering a redistribution of oxygen vacancies.*

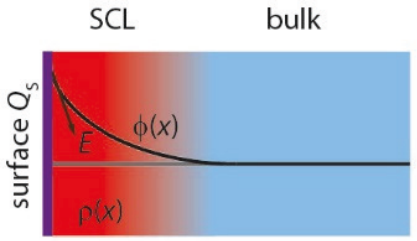


Fig. 17: *Schematic of the space charge potential established at the surface of acceptor-doped $SrTiO_3$. A surface charge Q_c generates a non-zero electric field at the surface, that is screened within the SCL.*

Hence, defects with a negative charge ($z_{def}<0$), such as electrons, will accumulate exponentially in the space charge layer, and defects with a positive charge ($z_{def}>0$), such as electron holes and oxygen vacancies will be depleted. The defect concentrations in space charge layers can differ by many orders of magnitude from their bulk values. The exponential behavior is stronger for doubly charged defects such as oxygen vacancies and strontium vacancies, so that these ionic defects typically form sharper and steeper concentration profiles than electrons and electron holes (Fig. 18).

Within the space charge layer *local charge neutrality* is violated. Therefore, the local charge neutrality condition has to be replaced by a *global charge neutrality* condition. Locally eq. (31) now reads

$$p(x)+2c_{V_O^{\cdot\cdot}}(x)-n(x)-c_{A'}=\rho(x), \tag{48}$$

where $\rho(x)$ denotes the local space charge density.

$\rho(x)$ determines the curvature of the potential via Poisson's equation

$$\frac{d^2\phi}{dx^2}(x)=\frac{\rho(x)}{\varepsilon_0\varepsilon_r}=\frac{1}{\varepsilon_0\varepsilon_r}\left[\sum_{def}c_{def}\exp\left(-\frac{z_{def}e\phi(x)}{k_BT}\right)-c_{A'}\right]. \tag{49}$$

Poisson's equation now explicitly depends on the potential itself via the local defect concentrations. (For weak accumulation of electrons, the constant acceptor concentration, $c_{A'}$, dominates $\rho(x)$ yielding the classical Mott-Schottky equation.)

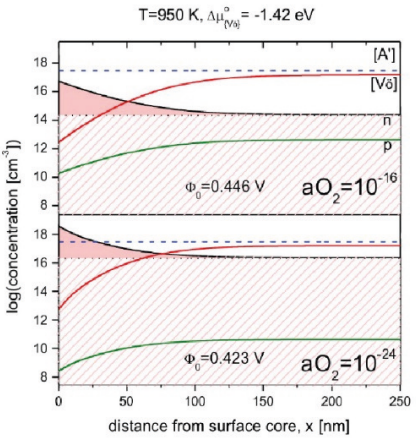


Fig. 18: *Calculated defect concentration profiles at the surface of $SrTiO_3$ thin films. Redistribution of oxygen vacancies towards the surface mediates space charge formation similar to the one at grain boundaries in ceramic compounds. In the SCL, n (black line) is by up to 2 orders of magnitude larger than in the bulk. The red shaded area indicates excess electrons as compared to the bulk concentrations.*

The global charge neutrality condition can be written in terms of Gauss' law

$$E\big|_{x=\infty} = \frac{Q\ (x=\infty)}{\varepsilon_0 \varepsilon_r} = \int_0^\infty \rho(x)dx = 0\,. \tag{50}$$

Hence, the total charge accumulated in core and space charge layer has to be zero, which implies that the electrical field is fully screened inside the bulk, and thus a flat potential.

For a given surface charge Q_c, eqs. (46)-(50) fully define the electrostatic problem which can be solved numerically in order to obtain the established potential.

As a result of space charge formation, *n*-type conduction along grain boundaries and surfaces of acceptor-doped $SrTiO_3$ is increased in comparison to the bulk. Moreover, it has been shown that the electronic conductivity of epitaxial thin films is enhanced compared to the bulk, because of the increasing weight of the conductance contribution of the surface space charge layer (Fig. 19) [19].

In contrast to that, one observes a depletion of electrons in the space charge layer at the surface of donor-doped $SrTiO_3$. Here, the surface charge is negative and presumably provided by the accumulation of strontium vacancies at the surface upon oxidation.

Also, 2-dimensional electron gases established at interfaces to $SrTiO_3$ may treated in terms of electrochemical space charge formation, triggering a mixed ionic-electronic interface reconstruction. [13]

8 Extended defect structures

So far, we focused on lattice disorder effects induced by point defects statistically distributed in the solid crystal. As we have seen, these homogeneously distributed point defects can be described in terms of thermodynamic equilibrium theories. However, also more complex defect clusters and extended defect structures can occur in real solids.

This is particularly the case for systems with high defect concentrations [20], which cannot be treated in a diluted manner any more. In such systems, individual defects are no longer independent from one another. Instead, they interact, potentially forming defect agglomerations,

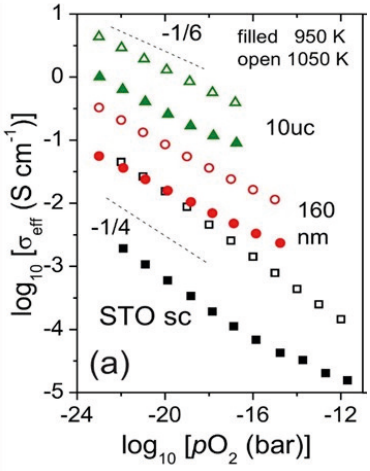


Fig. 19: *Experimentally determined conductivity of $SrTiO_3$ thin films in reducing atmosphere. The conductivity increases with decreasing layer thickness, because of the enhanced electronic conduction along the surface. This effect is based on space charge layer formation. Image taken from Ref. [19].*

such as defect complexes consisting of positive and negative point defects interacting via Coulomb interaction. This is observed in Ti-doped Nb_2O_5 where positively charged oxygen vacancies are attracted by the negatively charged dopant centers, Ti'_{Nb}, forming a defect complex

$$Ti'_{Nb} + V_O^{\bullet\bullet} \rightleftharpoons \left(Ti'_{Nb}\, V_O^{\bullet\bullet}\right)^{\bullet} . \tag{51}$$

Such complexes can to some extend be treated in a similar way as single point defects. However, their interaction has to be accommodated in an appropriate manner as described in Ref. [3].

Furthermore, solids generally consist defect structures of higher dimensionality, such as dislocations, stacking faults, etc. All these defects contribute to lattice disorder, and may affect the macroscopic and microscopic physical properties of solids.

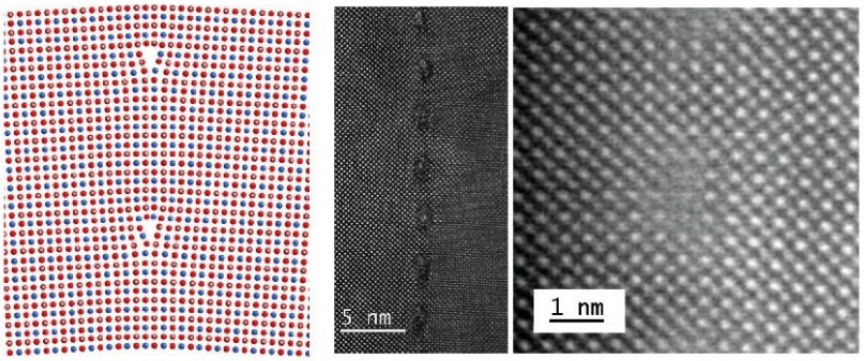


Fig. 20: *Equidistant dislocation cores along the boundary of a $SrTiO_3$ bi-crystal. (left) atomistic model; (right) high resolution STEM image (dark field) of a real bi-crystal boundary. Images adopted from Ref. [21].*

Extended defect structures can have various origin. Mostly, they are introduced during crystal growth or due to mechanical treatments (e.g. polishing), but also due to energetics. As an example, the boundary of a bi-crystal is decorated by equidistant dislocation cores (Fig. 20). The dislocations then accommodate the lattice mismatch caused by the different crystallographic orientations of the bi-crystal in the energetically most favorable way.

In the rutile phase of TiO_2, oxygen vacancies can arrange in ordered arrays when their concentration is high enough. In that case, the vacancies can be eliminated in a 'crystallographic shearing' process. The vacancy-rich layer is then replaced by a cation-rich layer, the 'shear planes', which appear as extra peaks in the x-ray diffraction pattern of the solid.

In the case of thin films, extended defect structures and even voids may be induced during thin film deposition. As shown for Sr_2TiO_4 and $SrTiO_3$ thin films, both the point defect structure as well as the dislocation and stacking fault density in the thin film can be controlled via growth parameters in a pulsed laser deposition experiment (Fig. 21). Depending on defect structure very different electronic behavior can be found, e.g. different resistive switching behavior [22].

Moreover, extended defects may arise as a result of phase separation in highly doped systems. Among these complex types of lattice disorder are the formation of Magnéli-phases (Ti_nO_{2n-1} [23]) and Wadsley defects in non-stoichiometric TiO_2. In Sr-rich $SrTiO_3$ crystals, often so-

called Ruddlesden-Popper phases ($Sr_{n+1}Ti_nO_{3n+1}$) are observed [24]. Moreover, SrO segregation, and thus phase separation, has been identified as aging effect in typical cathode materials used in solid oxide fuel cells, such as $(La,Sr)(Co,Fe)O_3$ and $(La,Ba)(Co,Fe)O_3$.

For a general discussion of the physical properties of ionic oxides both the point defect structure as well as the extended defect structure play an important role, while in the literature these two aspects are often treated in a separate manner.

As it turns out, it is often feasible to treat a system in a merely diluted point defect approach. A lot of physical properties and effects can be understood in such an approach. Even in locally confined systems, such as resistively switching memory cells, these models can be used to gain a qualitative understanding of microscopic phenomena and the involved thermodynamic driving forces. Nevertheless, on the local scale real thermodynamic processes and the defects structure in ionic materials may be more complex than indicated in this lecture.

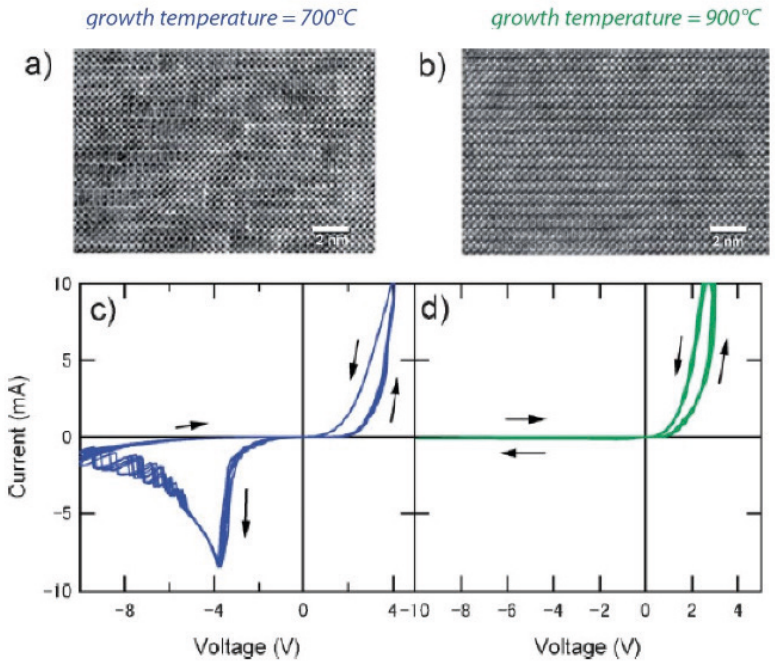


Fig. 21: *Defect structure of Sr_2TiO_4 thin films obtained by pulsed laser deposition at different growth temperature. While a high density of stacking faults (a) is established at 700°C, merely any extended defect structures are found for a growth temperature of 900°C (b). The two thin films show fundamentally different resistive switching behavior (c,d). Images adopted from Ref. [22].*

9 Amorphous materials

The topics of this Spring School are not limited to crystalline oxides or more generally to crystalline solids. In particular, phase change materials that will be covered in the remainder of this workshop can be intentionally driven into a non-crystalline, amorphous state. The controlled phase transition from the crystalline to the amorphous state (and vice versa) is then used to store information.

Here, we will give a short overview on the amorphous state of materials referring to a highly disordered, metastable state of matter. One particular issue for the discussion of amorphous systems is whether or not one can define *lattice disorder* in such highly *disordered* systems. A more detailed review on amorphous materials is given e.g. in Ref. [4].

While in crystalline solids the atoms (or ions) are arranged in a well-defined periodical manner, which implies long-range order, the atoms (or ions) are arranged in on randomly distributed sites (Fig. 22), which thus breaks translational symmetry of the crystalline phase. Hence, the amorphous state of a solid is a non-crystalline state, which *misses long-range order*.

Among the amorphous solids, the class of *glasses* are of great relevance. In glasses, the amorphous-crystalline phase transition can be controlled thermally. Typical glasses are SiO_2, $GeSe_2$ and the chalcogenides with great use in phase change memories.

In the random atomic pattern of glasses, the atomic distances (corresponding to the lattice constant in crystalline solids) varies locally (Fig. 22). Therefore, the amorphous state deviates from the state of minimum potential energy, which is – by definition – the crystalline phase. Therefore, amorphous materials are metastable and tend to crystalize. For this reason, it is not necessarily straight forward, how to generate an amorphous solid. Crystallization has to be suppressed kinetically in order to stabilize the disordered amorphous state.

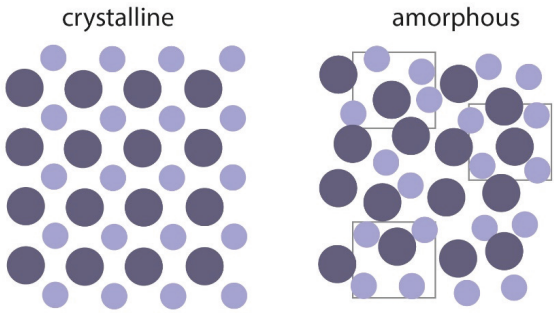


Fig. 22: *Crystalline vs. amorphous structure of a solid.*

The random distribution of atomic sites in amorphous solids reminds one of the atomic configuration of liquids and gases. Indeed, one important way of generating amorphous materials is quenching, i.e. fast cooling, of a liquid (melt-quenching) or gaseous phase of a material.

Upon cooling of a liquid, the material typically arranges in a crystalline manner – as far as the cooling procedure is slow enough to allow the atoms (or ions) to find their energetically favored positions in the lattice. However, if the cooling process is fast, the atomic arrangements freeze instantaneously resulting in supercooled liquids and ultimately in amorphous solids. Thus, whether or not an amorphous or crystalline state is achieved in a solid upon cooling strongly depends on the cooling rate. It is this effect, which is employed in phase change memory devices.

Quenching can for instance be established during condensation of a gaseous material on a cold substrate such as employed in chemical vapor deposition (CVD) techniques or plasma-based deposition techniques such as sputtering or pulsed laser deposition (at room temperature). Moreover, local heating e.g. by a laser can be used to locally (quasi-)melt a solid and to locally *write* an amorphous or crystalline state.

9.1 Defects in amorphous solids – concept of dangling bonds

The amorphous state of a solid is by definition a highly disordered one. Therefore, it is naturally difficult to define intrinsic lattice disorder for such a system or to define a (point) defect. For crystalline solids a defect is defined as a deviation from the perfectly ordered lattice. However, amorphous materials do not have a periodic lattice. Hence, how can one define a deviation from this, as a proper reference frame is missing?

While amorphous materials lack long-range order, and thus lack translational symmetry, they still exhibit forms of short-range and mid-range order. Locally, the entities still form bonds with their neighbors. Thus, in order to satisfy all bonds, the local coordination of the atoms (or ions) has to be similar to the crystalline phase. Therefore, also the stoichiometry is conserved locally (see boxes in Fig. 22). Thus, on the length scale of a few unit cells, amorphous materials possess next-neighbor order and a defined atomic coordination.

In this view, we can use deviations from this local short-range order to define an intrinsic point defect in an amorphous material. In particular, this can be done via the number of missing (or substituted) next neighbors. In most cases, this is equivalent to the number of *dangling bonds*, i.e. bonds that are non-saturated. Dangling bonds may therefore be used to define a (point) defect even in an amorphous material.

Such concepts can be used to understand qualitatively physical properties of amorphous materials such as resistivity, dielectric constant and diffraction index, in comparison to their crystalline counterparts.

10 Summary

In this lecture, the fundamental thermodynamic processes leading to lattice disorder in solids have been discussed. As shown, defects are a natural part of a solid and form as a result of minimization of Gibbs free energy. Defects are thus unavoidable in real systems. The defect structure of a solid, in particular of ionic oxides, can have huge impact on the physical properties of the material.

In the following lectures, lattice disorder, defect formation processes and phase transitions will be exploited to tailor material properties in a controlled way. To this end, defect structure can be used e.g. to store information which reflects the main topic of this Spring School. As indicated in the forgoing sections, however, the meaning of lattice disorder is much wider covering not only information storage but also energy conversion concepts, sensing devices, dielectrics, and novel transistor concepts. In all these applications and disciplines, lattice disorder has to be taken into account in order to understand the underlying physics as a whole. More detailed examples of lattice disorder effects will be given throughout the workshop.

References

[1] R. Waser (Ed.), *Nanoelectronics and Information Technology,* 3rd ed., Wiley-VCH (2012).

[2] H. Y. Hwang, Y. Iwasa, M. Kawasaki, B. Keimer, N. Nagaosa, and Y. Tokura, Emergent phenomena at oxide interfaces, *Nature Materials* **11**, 103–113 (2012).

[3] D. M. Smyth, *The Defect Chemistry of Metal Oxides*, Oxford University Press, Inc., New York (2000).

[4] C. R. A. Catlow (Ed.), *Defects and Disorder in Crystalline and Amorphous Solids*, Nato Science Series C, Vol. **418**, Springer Netherlands (1994).

[5] F. Kroger and H. Vink, *Relations Between the Concentrations of Imperfections in Crystalline Solids*, Solid State Physics-Advances in Reasearch and Applications **3**, 307-435 (1956).

[6] H. Ibach, H. Lüth, *Solid State Physics: An Introduction to Principles of Materials Science*, Springer Verlag, Berlin (2009).

[7] S. M. Sze, *Physics of Semiconductor Devices*, John Wiley & Sons, Inc., New York, (1981).

[8] R. Moos, K.H. Härdtl, "Defect Chemistry of Donor-Doped and Undoped $SrTiO_3$ Ceramics between 1000° and 1400°C", *J. Am. Ceram. Soc.* **80**, 2549-62 (1997).

[9] Y. Kozuka, M. Kim, C. Bell, B. G. Kim, Y. Hikita & H. Y. Hwang, Two-dimensional normal-state quantum oscillations in a superconducting heterostructure, *Nature* **462**, 487-490 (2009).

[10] R. Merkle, J. Maier, "How Is Oxygen Incorporated into Oxides? A Comprehensive Kinetic Study of a Simple Solid-State Reaction with $SrTiO_3$ as a Model Material", *Angew. Chem. Int. Ed.,* **47**, 3874-3894 (2008).

[11] F. Gunkel, The role of defects at functional interfaces between polar and non-polar perovskite oxides, *dissertation*, RWTH Aachen, Aachen, Germany (2013).

[12] F. Gunkel, S. Hoffmann-Eifert, R. Dittmann, S. Mi, C. Jia, P. Meuffels, and R. Waser, High temperature conductance characteristics of $LaAlO_3/SrTiO_3$-heterostructures under equilibrium oxygen atmospheres, Applied Physics Letters **97**, 12103 (2010).

[13] F. Gunkel, S.Wicklein, S. Hoffmann-Eifert, P. Meuffels, P. Brinks, M. Huijben, G. Rijnders, R. Waser, and R. Dittmann, "Transport limits in defect-engineered $LaAlO_3/SrTiO_3$ bilayers, Nanoscale **7**, 1013–1022 (2015).

[14] F. Gunkel, P. Brinks, S. Hoffmann-Eifert, R. Dittmann, M. Huijben, J. E.Kleibeuker, G. Koster, G. Rijnders, and R. Waser, Influence of charge compensation mechanisms on the sheet electron density at conducting $LaAlO_3/SrTiO_3$-interfaces," Applied Physics Letters **100**, 052103 (2012).

[15] M. Kubicek, A. Limbeck, T. Frömling, H. Hutter, and J. Fleig, Relationship between Cation Segregation and the Electrochemical Oxygen Reduction Kinetics of $La_{0.6}Sr_{0.4}CoO_{3-\delta}$ Thin Film Electrodes, *Journal of The Electrochemical Society* **158**, 6 (2011).

[16] R. Waser and R. Hagenbeck, Grain Boundaries in Dielectric and Mixed- Conducting Ceramics, *Acta Materialia* **48**, 797 (2000).

[17] R. A. De Souza, The formation of equilibrium space-charge zones at grain boundaries in the perovskite oxide $SrTiO_3$, *Physical Chemistry Chemical Physics* **11**, 43 (2009).

[18] R. A. De Souza, V. Metlenko, D. Park, and T. E. Weirich, Behavior of oxygen vacancies in single-crystal $SrTiO_3$: Equilibrium distribution and diffusion kinetics, *Physical Review B* **85**, 17 (2012).

[19] R. A. De Souza, F. Gunkel, S. Hoffmann-Eifert, and R. Dittmann, Finite-size versus interface-proximity effects in thin-film epitaxial $SrTiO_3$, *Physical Review B* **89**, 241401(R) (2014).

[20] K. Szot, W. Speier, R. Carius, U. Zastrow, W. Beyer, "Localized Metallic Conductivity and Self-Healing during Thermal Reduction of $SrTiO_3$", *Phys. Rev. Lett.* **88**, (2002).

[21] V. Metlenko, A. H. H. Ramadan, F. Gunkel, H. Du, H. Schraknepper, S. Hoffmann-Eifert, R. Dittmann, R. Waser, and R. A. De Souza, *Nanoscale* **6**, 12864 (2014).

[22] K. Shibuya, R. Dittmann, S. Mi, R. Waser, Impact of defect distribution on resistive switching characteristics of Sr_2TiO_4 thin films, *Advanced Materials* **22**, 3 (2010).

[23] S. Andersson, A. Sundholm, A. Magneli, *Acta Chem. Scand.* **13**, 989-997 (1959).

[24] S.N. Ruddlesden and P. Popper, "The compound Sr_3TiO_7 and its structure", *Acta Cryst.* **11**, 54 (1958).

A 4 Ion Transport in Metal Oxides

Roger A. De Souza
Institut für Physikalische Chemie
RWTH Aachen University, 52056 Aachen

Contents

1 Introduction **2**

2 Macroscopic Definition **2**
2.1 Two Solutions of the Diffusion Equation 3
2.2 Dependence of the diffusion coefficient on characteristic thermodynamic parameters 5

3 Microscopic Definition **5**
3.1 Mechanisms of diffusion 6
3.2 Diffusion coefficients of defects and ions 7
3.3 The activation barrier for migration 8

4 Types of Diffusion Experiments **10**
4.1 Chemical diffusion 11
4.2 Tracer diffusion 12
4.3 Conductivity 14

5 Mass transport along and across extended defects **16**
5.1 Accelerated transport *along* extended defects 18
5.2 Hindered transport *across* extended defects 20

Lecture Notes of the 47$^{\text{th}}$ IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

The subject of ion transport in metal oxides is both broad and deep, and accordingly there are many different ways of approaching the subject. This is due in part to the variety of driving forces that can cause ions to move. For example, ion motion may occur because of a concentration gradient, in which case one speaks of diffusion; or because of an electrical potential gradient, in which case one speaks of drift; or because of a temperature gradient, in which case one speaks of thermodiffusion. In this chapter we use diffusion in the solid state as a means of examining ion transport in oxides. After introductory sections covering the basics of diffusion, we consider the differences and the relationships between the most common diffusion coefficients. The possibilities of accelerated diffusion occurring along extended defects and diffusion across extended defects being hindered are also discussed. It is conceivable that, compared with diffusion in the bulk, diffusion along extended defects is retarded or that diffusion across extended defects is accelerated; these cases are of minor importance and, therefore, are not considered here. The final section provides a literature survey of anion and cation transport in several example systems.

There are two complementary approaches to treating diffusion: the macroscopic approach based on Fick's empirical equations; and the microscopic approach based on atomic mechanisms and on random-walk theory. The macroscopic definition provides the experimental basis for determining diffusion coefficients. At no point in the macroscopic treatment, however, does the atomic nature of matter enter into the treatment. Nonetheless, diffusion in solids results from many individual jumps of the diffusing particles, and it is mediated by point defects, such as vacancies and interstitials. The benefit of the microscopic approach, then, is that it provides the theoretical basis for interpreting and expressing diffusion coefficients in terms of atomic quantities, such as point-defect concentrations, atomic jump distances and activation barriers.

In writing this chapter, which has already appeared elsewhere[1, 2] I have referred to several books [3–8] and review articles [9–16]. My aim is to present the subject matter so that newcomers will be able in the end to understand, to apply, and perhaps even, to judge critically literature data for ion transport in metal oxides.

2 Macroscopic Definition

Let us consider, at the level of a continuum, a single-phase system, in which the chemical component i is distributed inhomogeneously (that is, its concentration c_i is not the same everywhere in the system). We now bring the system to a temperature at which i is mobile: As a result of diffusion, component i will move in a manner to decrease its concentration gradient. The flux j_i (that is, the amount of i passing through unit area of a reference plane per unit time) is given by Fick's first law. In one dimension (x), this law can be written as

$$j_i = -D_i \frac{\partial c_i}{\partial x}. \tag{1}$$

The factor of proportionality between the flux and the negative concentration gradient is, D_i, the diffusion coefficient of i in the specific system. The minus sign in Eq. (1) indicates that the flux is directed towards lower concentrations, *i.e.* 'down the concentration gradient'. As expected, Eq. (1) indicates that the flux goes to zero as $\partial c_i/\partial x$ goes to zero: the process of diffusion leads to the elimination of concentration gradients [3, 4].

In the majority of cases, the amount of the diffusing component is conserved during the process of diffusion; this means that component i is neither created nor destroyed. If we consider a small volume element within our continuum phase, the flux of component i into this small volume element minus the flux of i flowing out of it equals the rate of accumulation (or loss) of i within that volume element. This statement is captured mathematically by the continuity equation, which for diffusion in one dimension is

$$-\frac{\partial j_i}{\partial x} = \frac{\partial c_i}{\partial t}. \tag{2}$$

Substitution of Eq. (1) into Eq. (2) yields Fick's Second Law,

$$\frac{\partial c_i}{\partial t} = \frac{\partial}{\partial x}\left(D_i \frac{\partial c_i}{\partial x}\right), \tag{3}$$

which, if the diffusion coefficient is independent of position, reduces to

$$\frac{\partial c_i}{\partial t} = D_i \frac{\partial^2 c_i}{\partial x^2}. \tag{4}$$

The second-order partial differential equation, Eq. (3) [or (4)], is sometimes called the 'diffusion equation'.

Although the one-dimensional treatment is often sufficient, Fick's first law is easily generalised to the three-dimensional case:

$$\mathbf{j}_i = -\mathbf{D}_i \nabla c_i \tag{5}$$

in which $\mathbf{j}_i$ is the vector flux; the Nabla operator ∇ acts on the scalar concentration field $c_i(x, y, z)$; and $\mathbf{D}_i$ is a symmetric second-rank tensor (in cubic crystals, it reduces to the single scalar quantity D_i). Analogous phenomena, in the sense of cause (driving force) and effect (flux) are Fouriers law of heat transport, $j_Q = -\kappa \nabla T$ (a flux of heat, j_Q, is driven by a gradient in temperature T, with the thermal conductivity κ as the proportionality constant), and Ohms law, $I = -\sigma \nabla \phi$ (the current density I is driven by a gradient in electrical potential ϕ, with the electrical conductivity σ as the proportionality constant) [4].

It is important to note that, although Eq. (1) gives the macroscopic definition of a diffusion coefficient, the true driving force for diffusion is not the gradient in the concentration of i (∇c_i), but the gradient in its chemical potential ($\nabla \mu_i$); see Section 4. At chemical equilibrium, the gradient in the chemical potential of all mobile components is zero.

2.1 Two Solutions of the Diffusion Equation

From a practical point-of-view, Fick's First Law [Eq. (1)] is not particularly useful for diffusion measurements in the solid state because it requires us to determine the diffusion flux. The accurate measurement of fluxes is seldom trivial, and often virtually impossible; in addition, solid-state diffusivities are often small, so that the corresponding fluxes are also small. The optimal basis for determining a diffusion coefficient has proven to be the examination of concentration transients. One measures, for example, the concentration profile in a diffusion specimen as a function of position for a given time, and the diffusion coefficient D_i is obtained by describing the concentration profile with an appropriate solution to the diffusion equation [Eq. (4 or 5)] for the given initial and boundary conditions.

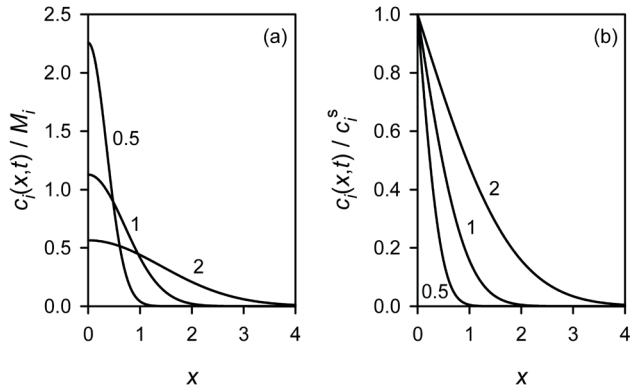


Fig. 1: *Solutions of the diffusion equation for the case of an instantaneous source (a) and constant source (b) plotted as normalised concentration against depth x for various values of $\sqrt{4D_it}$. The units of x and $\sqrt{4D_it}$ are arbitrary but identical (if x is expressed in cm, D_i has units of cm^2 s^{-1} and t has units of s).*

One point cannot be emphasised enough: If one can describe an entire concentration profile with the appropriate solution of the diffusion equation, one has incontrovertible evidence that diffusion in the solid state has taken place.

For the sake of illustration, we consider two experimental arrangements that are often used in diffusion studies. In both cases, the experiment is set up so that diffusion takes place in one dimension (x), in a medium that extends to infinity in the positive x direction ($x > 0$, semi-infinite medium) and that contains no diffusant at the start of the experiment, $c_i(x > 0, t = 0) = 0$; furthermore the diffusion coefficient D_i does not depend on position, so that Eq. (4) is the differential equation to be solved. The two cases we consider are referred to as the instantaneous source solution and the constant source solution (Fig. 1). For a comprehensive treatment of the mathematics of diffusion, the reader is referred to the textbook of Crank [17].

In the first case, a small amount of diffusant is deposited at the plane $x = 0$. Its initial distribution is given by

$$c_i(x, 0) = M_i\delta(x), \tag{6}$$

M_i denotes the amount of diffusant per unit area and $\delta(x)$ the Dirac delta function. After diffusion for a time t the concentration profile of i in the sample is described by

$$c_i(x, t) = \frac{M_i}{\sqrt{\pi D_i t}} \exp\left(-\frac{x^2}{4D_i t}\right). \tag{7}$$

The quantity $\sqrt{4D_it}$ is a characteristic diffusion length.

In the second experimental arrangement, the surface of a semi-infinite medium is exposed continuously to a fixed concentration of diffusant (c_i^s), for example, by exposing the surface to an atmosphere of the diffusant. The corresponding solution of the diffusion equation is

$$c_i(x, t) = c_i^s \operatorname{erfc}\left(\frac{x}{2\sqrt{D_i t}}\right) \tag{8}$$

where $\mathrm{erfc}(z)$ is the complementary error function: $\mathrm{erfc}(z) = 1 - \mathrm{erf}(z)$, with $\mathrm{erf}(z)$ being the Gaussian error function (a standard mathematical function), defined as $\mathrm{erf}(z) \equiv \frac{2}{\sqrt{\pi}} \int_0^z \mathrm{e}^{-u^2} \mathrm{d}u$.

2.2 Dependence of the diffusion coefficient on characteristic thermodynamic parameters

Phenomenologically one often observes that, as a function of temperature T, the diffusion coefficient obeys Arrhenius type behaviour,

$$D_i(T) = D_{i,0} \exp\left(-\frac{\Delta H_{D_i}}{k_\mathrm{B} T}\right), \tag{9}$$

that is, the behaviour is characterized by just two quantities, the pre-exponential factor $D_{i,0}$ and the activation enthalpy of diffusion ΔH_{D_i}. Values of ΔH_{D_i} determined experimentally may be as low as some tenths of an eV or as high as 10 eV, but are generally of the order of 10^0 eV.
For a binary metal oxide MO, the Gibbs phase rule stipulates that three variables are required to define the system thermodynamically. Two variables are temperature and hydrostatic pressure; the third variable is commonly chosen to be the oxygen partial pressure $p\mathrm{O}_2$ (as it is easier experimentally to define and to vary $p\mathrm{O}_2$ than pM, the partial pressure of the metallic component; furthermore $p\mathrm{O}_2$ can be varied experimentally over tens of orders of magnitude). Values of the diffusion coefficient measured as a function of $p\mathrm{O}_2$ at constant temperature are found to obey a power law, with the power-law exponents $m_{p\mathrm{O}_2}$,

$$m_{p\mathrm{O}_2} = \left(\frac{\partial \ln D_i}{\partial \ln p\mathrm{O}_2}\right)_T. \tag{10}$$

Considered over a suitably large range of oxygen partial pressures, $m_{p\mathrm{O}_2}$ may take characteristic values of $\pm\frac{1}{6}$, $\pm\frac{1}{4}$, $\pm\frac{1}{2}$, *etc.* (see chapter by F. Gunkel).
Similarly, for a doped binary oxide, at defined T and $p\mathrm{O}_2$, one often observes that D_i varies with dopant concentration $[\mathrm{Dop}]$ according to a power law, with exponent

$$m_{[\mathrm{Dop}]} = \left(\frac{\partial \ln D_i}{\partial \ln[\mathrm{Dop}]}\right)_{T,p\mathrm{O}_2}. \tag{11}$$

In order to interpret and understand the activation energy of diffusion or one of the power-law exponents, one must consider diffusion from a microscopic standpoint.

3 Microscopic Definition

Fick's first law, as described in Eq. (1) or (5), provides the macroscopic definition of the diffusion coefficient D_i. The microscopic definition comes, independently, from Einstein and from Smoluchowski. They considered that the diffusion arises as a result of the motion of atomic species (atoms, ions, molecules), and they examined the question of how far these species diffuse in a given time. They recognised that the quantity of interest is not the velocity of the diffusing species, but their mean square displacement. Specifically they related the mean square displacement $\langle x^2 \rangle$ of the diffusing species in the x-direction to the x-component of the diffusion coefficient, D_i^x, and to the time interval t during which diffusion takes place [3, 4, 7]:

$$\left\langle x_i^2 \right\rangle = 2 D_i^x t. \tag{12}$$

This is now known as the Einstein relation or as the Einstein–Smoluchowski relation. For diffusion in an isotropic, three-dimensional medium, $\langle x_i^2 \rangle = \langle y_i^2 \rangle = \langle z_i^2 \rangle = \langle R_i^2 \rangle /3$, the mean square displacement is given by

$$\left\langle R_i^2 \right\rangle = 6 D_i t. \tag{13}$$

In the microscopic picture, diffusion is considered to occur as the net result of many individual atomic jumps. Let us consider, then, a single particle (atom, ion, molecule) diffusing on an otherwise empty, three-dimensional lattice. The particle executes a series of consecutive jumps, each of distance a, from one site to a neighbouring site (a in oxides is thus of the order of a few Å). Having made n jumps, the particle is characterised (Fig. 2) by a displacement R from its starting position, and a squared displacement R^2. If each jump is independent of the previous jumps, that is, the particle has no memory of past jumps made, the particle is said to make an uncorrelated random walk. Let us now repeat the walk of n jumps a large number of times and consider the average quantities. (As an alternative to averaging repeated trials of an isolated particle, one can average over a large set of dilute, non-interacting particles.) The average displacement $\langle R_i \rangle$ in this case is zero, because on average the number of jumps in $+x$, $-x$, *etc.* will be the same. The mean square displacement $\langle R_i^2 \rangle$, however, is not zero. From the theory of random walks, one obtains for an uncorrelated random walker

$$\left\langle R_i^2 \right\rangle = n a^2. \tag{14}$$

Combining Eqs. (13) and (14), and introducing the jump rate of the particle to one of its Z neighbours, $\Gamma_i Z = n/t$, one obtains

$$D_i = \frac{1}{6} a^2 Z \Gamma_i. \tag{15}$$

That is, we have expressed the diffusion coefficient in terms of certain atomic quantities: the particle's jump distance to a neighbouring site a, the number of its neighbouring sites Z and its jump rate Γ_i [4].

3.1 Mechanisms of diffusion

A variety of mechanisms have been proposed over the last century to explain atomic motion in crystalline solids. Nowadays there is ample evidence that diffusion in crystalline solids takes place by defect-mediated mechanisms. Three common mechanisms are depicted in Fig. 3.

In the interstitial mechanism (a), sometimes called the direct interstitial mechanism, an interstitial ion executes a jump from one interstice to another. In the indirect interstitial mechanism (b), known more commonly as the interstitialcy mechanism, an ion on an interstitial lattice site pushes an ion on a regular lattice site onto an interstitial site and occupies the regular lattice site for itself. This process can occur with the three species moving in a direct line (collinear) or at an angle to one another (non-collinear). The third mechanism, and probably the most important, is the vacancy mechanism (c): an ion occupying a regular lattice site jumps to a neighbouring unoccupied site, that is, the migrating ion and the vacancy exchange sites.

For the sake of completeness we also mention two mechanisms that were believed for a long time to be operative in crystalline solids, but are now best regarded as hypothetical possibilities. The direct exchange of neighbouring ions, in which two ions move simultaneously, requires large distortions of the lattice as the migrating species squeeze past each other; this makes this process energetically improbable for ionic solids. The ring mechanism, corresponding to the rotation of 3 (or more) ions as a group, requires less lattice distortion, but the collective nature of this complex mechanism makes it unlikely for most crystalline substances.

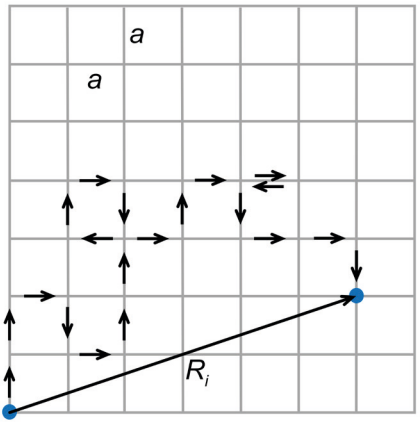


Fig. 2: *A single random walk of a single particle i on an empty two-dimensional lattice, with intersite jump distance of a. After $n = 20$ jumps the particle has achieved a displacement R_i. Repeating the random walk of $n = 20$ jumps many times yields $\langle R_i \rangle = 0$ but $\langle R_i^2 \rangle = 20a^2$. Note: for each random walk the particle covers a distance of $20a$; after many random walks, however, its root-mean-square displacement is $\sqrt{20}a$. For this two-dimensional case the diffusion coefficient follows as $D_i = \frac{1}{4}a^2 Z\Gamma_i$, with $Z = 4$. Adapted from [4].*

3.2 Diffusion coefficients of defects and ions

In general there will be a considerable difference between the diffusion coefficient of the defect, D_{def}, and the diffusion of the ion, D_{ion}. To see why this is so, we consider the concrete example of oxygen ions migrating in an oxide by a vacancy mechanism. The diffusion coefficient of the vacancies is [see Eq. (15)]

$$D_{\mathrm{V}} = \frac{1}{6}a^2 Z\Gamma_{\mathrm{V}}, \tag{16}$$

whereas the diffusion coefficient of the oxygen ions is

$$D_{\mathrm{O}} = \frac{1}{6}a^2 Z\Gamma_{\mathrm{O}}. \tag{17}$$

Each time a vacancy moves, an ion has to move, since the two species swap places. Thus, the total number of displacements of the ions and of the vacancies has to be equal

$$\Gamma_{\mathrm{O}}[\mathrm{O}] = \Gamma_{\mathrm{V}}[\mathrm{V}], \tag{18}$$

with $[i]$ denoting the concentration of i. For a dilute solution of vacancies, the site fraction of vacancies, n_{V}, is by definition many orders of magnitude less than unity, and it can thus be approximated by $n_{\mathrm{V}} \approx [\mathrm{V}]/[\mathrm{O}]$. Hence, by combining Eqs. (16), (17) and (18), we obtain

$$D_{\mathrm{O}} = D_{\mathrm{V}}\frac{[\mathrm{V}]}{[\mathrm{O}]} \approx D_{\mathrm{V}} n_{\mathrm{V}}. \tag{19}$$

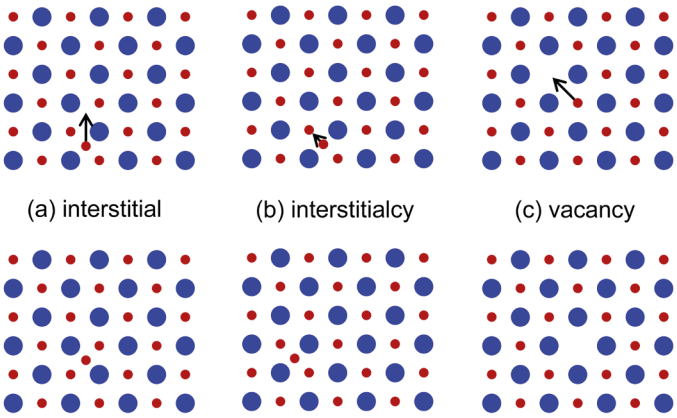


Fig. 3: *Common mechanisms of atomic migration in a binary compound with mobile species (red) and immobile species (blue): (a) interstitial (b) interstitialcy (c) vacancy.*

Thus, D_{O} and D_{V} will differ by the factor $n_{\mathrm{V}} \ll 1$. One should note that n_{V} may vary considerably with temperature, oxygen partial pressure and dopant concentration (see chapter by F. Gunkel). D_{V}, at least for dilute solutions of vacancies, is independent of n_{V}, and as we will see later, only dependent on temperature. One consequence of Eq. (19) is that the mean square displacement of a vacancy, $\langle R_{\mathrm{V}}^2 \rangle$, will be orders of magnitude larger than the mean square displacement of an oxygen ion, $\langle R_{\mathrm{O}}^2 \rangle$.
A similar expression to Eq. (18) also holds for the diffusion of oxygen interstitials (I) in an oxide. Consequently, in an oxide with anti-Frenkel disorder, not only the oxygen vacancies but also the oxygen interstitials may contribute to the overall diffusion coefficient of the ions,

$$D_{\mathrm{O}}[\mathrm{O}] = D_{\mathrm{V}}[\mathrm{V}] + D_{\mathrm{I}}[\mathrm{I}]. \tag{20}$$

Eq. (19) can also be derived by noting that the probability of an oxygen-ion jump is equal to the product of the vacancy hopping rate, Γ_{V}, and the probability of finding a vacancy adjacent to the ion, n_{V}. Hence, in the case of impurity diffusion by an interstitial mechanism (that is, the impurity interstitials remain on the interstitial sublattice and do not undergo exchange with regular lattice species), $D_{\mathrm{ion}} = D_{\mathrm{def}}$ because the probability of finding an vacant interstitial site is essentially unity (for a dilute solution).

3.3 The activation barrier for migration

In jumping from one site to another, an ion has to overcome an activation barrier, regardless of which of these three migration mechanisms is operative. Most of the time, the ion is vibrating around its equilibrium position (which may be a regular lattice site or an interstitial position). Occasionally, it successfully executes a jump to a new site. Thereafter it vibrates around its new position, waiting until it successfully makes the next jump. (As the jumps are random, the next successful jump may bring the ion back to its original site). In other words, the duration of a jump is short compared with the residence time of the ion on its site [4].

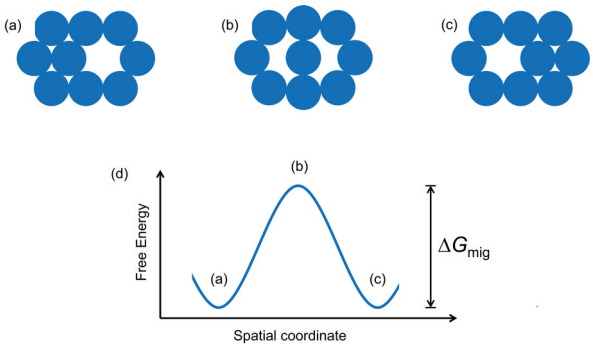


Fig. 4: *Simple schematic illustration of the ion movements involved in the jump of an ion to a neighbouring vacant site (migration mediated by a vacancy mechanism). (a) Initial configuration; (b) Saddle-point configuration; (c) Final configuration. (d) The change in the system's Gibbs energy associated with the ion movements in (a)-(c).* ΔG_{mig} *is the Gibbs activation energy of migration.*

In Fig. 4(a)-(c) we see a simple schematic illustration of the migration process for the case of a vacancy mechanism. What is required for the ion to make a successful jump? In this simple picture, not only does the migrating ion have to move in the right direction, the surrounding lattice ions have to move out of the way. Only if both occur will the ion complete its jump. Fig. 4 (d) shows the associated change in the Gibbs energy of the system as the migrating ion is moved reversibly from (a) to (c) [3]; ΔG_{mig}, the Gibbs activation energy of migration, corresponds to the difference in the system's Gibbs energy between the initial and saddle-point configurations. The jump rate Γ is related to ΔG_{mig} through an an Arrhenius law,

$$\Gamma = \nu_0 \exp\left(\frac{-\Delta G_{\mathrm{mig}}}{k_{\mathrm{B}} T}\right), \tag{21}$$

where the prefactor ν_0 denotes an attempt frequency of the order of the Debye frequency of the compound, k_{B} is Boltzmann's constant, and T is the absolute temperature. This is a simple description of a complicated many-body problem. For the detailed consideration of defect migration in solids including many-body effects, the interested reader is referred to Vineyard [18]. ΔG_{mig}, being a Gibbs energy, can be expressed in terms of an activation enthalpy of migration, ΔH_{mig}, and an activation entropy of migration, ΔS_{mig},

$$\Delta G_{\mathrm{mig}} = \Delta H_{\mathrm{mig}} - T\Delta S_{\mathrm{mig}}. \tag{22}$$

Combining Eqs. (19), (21) and (22), we thus find, for the case of oxygen-ion diffusion mediated by a vacancy mechanism in a cubic lattice, that the self-diffusion coefficient of oxygen is given by

$$D_{\mathrm{O}} = n_{\mathrm{V}}(T, p\mathrm{O}_2)\frac{1}{6}a^2 Z\nu_0 \exp\left(\frac{\Delta S_{\mathrm{mig,V}}}{k_{\mathrm{B}}}\right)\exp\left(\frac{-\Delta H_{\mathrm{mig,V}}}{k_{\mathrm{B}} T}\right). \tag{23}$$

4 Types of Diffusion Experiments

There are a bewildering variety of ways in which to carry out diffusion experiments, and thus, a bewildering variety of diffusion coefficients. For metal oxides, the three most commonly determined coefficients are D^*, the tracer diffusion coefficient; D^δ, the chemical diffusion coefficient; and D^σ, the conductivity diffusion coefficient [8]. In the following we will examine, first, what the basics of the three diffusion experiments are, and second, how the respective diffusion coefficients are related to the self-diffusion coefficients of the ions D_{ion}, or of the defects D_{def}.

In order to obtain expressions relating the diffusion coefficients to one another, we require a theoretical framework to treat these processes, and the most convenient theoretical framework for treating these and other transport processes is linear irreversible thermodynamics [19]. For example, for a system that contains the mobile charge carriers (ions, electrons) i and k and that is exposed to a temperature gradient, one can express, according to linear irreversible thermodynamics, the vector fluxes of the two charge carriers and the vector flux of heat as

$$\mathbf{j}_i = L_{ii}\mathbf{X}_i + L_{ik}\mathbf{X}_k + L_{iQ}\mathbf{X}_Q \tag{24a}$$

$$\mathbf{j}_k = L_{ki}\mathbf{X}_i + L_{kk}\mathbf{X}_k + L_{kQ}\mathbf{X}_Q \tag{24b}$$

$$\mathbf{j}_Q = L_{Qi}\mathbf{X}_i + L_{Qk}\mathbf{X}_k + L_{QQ}\mathbf{X}_Q \tag{24c}$$

that is, in terms of phenomenological transport coefficients L and general thermodynamic driving forces, $\mathbf{X}$. Eq. (24) states that a flux of i, say, can result directly from the driving force $\mathbf{X}_i$, but also indirectly from the driving forces $\mathbf{X}_k$ and $\mathbf{X}_Q$. The thermodynamic driving force for heat transport is given by $\mathbf{X}_Q = -(1/T)\nabla T$; for the charge carrier i, the thermodynamic driving force is given, in the absence of external forces, by $\mathbf{X}_i = -T\nabla(\eta_i/T)$, where η_i denotes the electrochemical potential of i and is defined as (z_i is the charge number of i)

$$\eta_i = \mu_i + z_i e\phi. \tag{25}$$

The off-diagonal elements of the matrix of L coefficients lead to 'cross' effects, such as thermoelectricity (L_{eQ}) and thermodiffusion (L_{iQ}) and are thus also known as cross-coefficients. Onsager's reciprocity theorem states that, in the absence of a magnetic field, the matrix of L coefficients is symmetric, that is, $L_{ik} = L_{ki}$, *etc.* It is noted that an oxide was used as a model system to verify Onsager's theorem experimentally [20].

Let us now for simplicity restrict the treatment to the one dimensional case and to ideal solutions. Transforming Eq. (24) to the 'reduced' heat formulation [9, 15, 16, 21]; assuming $L_{ik} = L_{ki}$, *etc.*; and using $L_{ii} = D_i[i]/k_{\mathrm{B}}T$; one can write the flux of i as

$$j_i = -\frac{D_i[i]}{k_{\mathrm{B}}T}\left[\frac{\partial \eta_i}{\partial x} + \left(\bar{S}_i + \frac{Q_i}{T}\right)\frac{\partial T}{\partial x}\right]. \tag{26}$$

where $\bar{S}_i$ is the partial molar entropy and Q_i the reduced heat of transport of charge carrier i. Thus we perceive that a flux of i can be driven by a gradient in chemical potential (diffusion), by a gradient in electrical potential (drift) and by a gradient in temperature (thermodiffusion). The term $(\bar{S}_i + Q_i/T)$ may be positive or negative, and thus depending on the sign, the ion i may move down or up the temperature gradient. In the isothermal case, to which the three diffusion experiments refer, Eq. (26) reduces to

$$j_i = -\frac{D_i[i]}{k_{\mathrm{B}}T}\left[\frac{\partial \mu_i}{\partial x} + z_i e\frac{\partial \phi}{\partial x}\right]. \tag{27}$$

The procedures for deriving the various diffusion coefficients from Eq. (24) or from from Eq. (27) are given in detail elsewhere [9, 10, 13]; for reasons of limited space, only the end results are reproduced in the relevant sections below. Furthermore, only the simplest cases are considered: many complications are possible [8, 13, 16, 22].

4.1 Chemical diffusion

Consider an oxide of composition $MO_{1-\delta}$ in chemical equilibrium with the surrounding gas phase; that is, the chemical potential of oxygen throughout the oxide and equal to that in the gas phase. A difference in the chemical potential of oxygen between the sample and the surrounding gas phase is now effected (either by changing the system's temperature or the oxygen activity of the gas phase), with the result that oxygen is either incorporated or removed from the oxide. The nonstoichiometry changes from its original value δ to a final value $\delta + \Delta\delta$. The process whereby this takes place ($MO_{1-\delta} + \Delta\delta \cdot \frac{1}{2}O_2 \rightarrow MO_{1-\delta+\Delta\delta}$) is called chemical diffusion. The chemical diffusion coefficient, according to Fick's first law [Eq. (1)], is denoted by D_i^δ, $\tilde{D}_i$ or $D_{i,\mathrm{chem}}$: The true driving force, however, is the gradient in the chemical potential of oxygen [13].

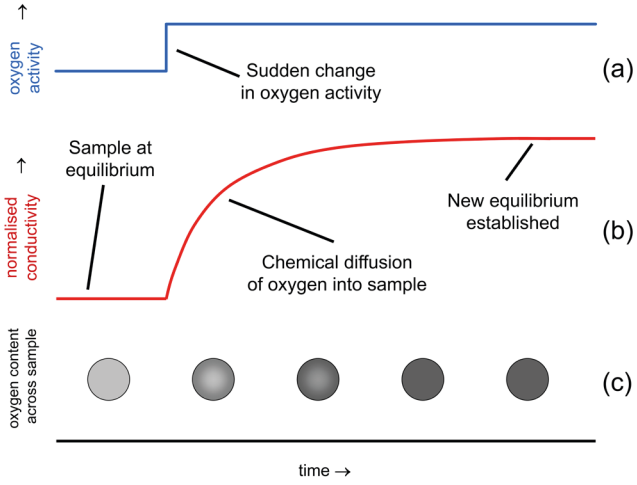


Fig. 5: *The chemical diffusion experiment consist, for example, of raising instantaneously the activity of oxygen in the surrounding gas phase (a), and then monitoring as a function of time the change in a characteristic sample property (b), i.e. one that depends of the oxygen content of the sample (c), such as the electrical conductivity, as the sample attains the new equilibrium with the gas phase.*

Although formally oxygen is incorporated into the oxide as a neutral component, at the microscopic scale there is coupled transport of charged species, otherwise known as ambipolar diffusion. Let us assume that our oxide $MO_{1-\delta}$ has oxygen vacancies and electrons as the dominant mobile defects. As oxygen is incorporated, there will be a flux of vacancies j_V and a flux

of electrons j_e towards the surface, these two fluxes being given by [see Eq. (27)]

$$j_\mathrm{V} = -\frac{D_\mathrm{V}[\mathrm{V}]}{k_\mathrm{B}T}\left[\frac{\partial \mu_\mathrm{V}}{\partial x} + 2e\frac{\partial \phi}{\partial x}\right] \tag{28a}$$

$$j_\mathrm{e} = -\frac{D_\mathrm{e}[\mathrm{e}]}{k_\mathrm{B}T}\left[\frac{\partial \mu_\mathrm{e}}{\partial x} - e\frac{\partial \phi}{\partial x}\right] \tag{28b}$$

These two fluxes are not independent, however, as $2j_\mathrm{V} - j_\mathrm{e} = 0$. Furthermore, the internal gradient in the electrical potential, $-\partial\phi/\partial x$ (also known as the Nernst field), is common to both fluxes; this coupling has the effect of accelerating the slower moving moiety and slowing down the faster moving one. The detailed analysis yields

$$D^\delta_\mathrm{O} = \frac{D_\mathrm{V}[\mathrm{V}]\, D_\mathrm{e}[\mathrm{e}]}{4D_\mathrm{V}[\mathrm{V}] + D_\mathrm{e}[\mathrm{e}]}\left(\frac{1}{2}\frac{\partial \ln p\mathrm{O}_2}{\partial c_\mathrm{O}}\right). \tag{29}$$

It is clear from Eq. (29) that the chemical diffusion coefficient of oxygen in $\mathrm{MO}_{1-\delta}$, D^δ_O, may exhibit rather complex behaviour as a function of temperature and oxygen partial pressure, depending on how the individual parameters vary with T and $p\mathrm{O}_2$.
In some cases, Eq. (29) reduces to a simpler form: If, for instance, the two defects are dilute, non-interacting species, we find

$$D^\delta_\mathrm{O} = \frac{D_\mathrm{V}[\mathrm{V}]\, D_\mathrm{e}[\mathrm{e}]}{4D_\mathrm{V}[\mathrm{V}] + D_\mathrm{e}[\mathrm{e}]}\left(\frac{1}{[\mathrm{V}]} + \frac{4}{[\mathrm{e}]}\right). \tag{30}$$

Furthermore, if the electrons are more mobile than the vacancies, $D_\mathrm{e} \gg D_\mathrm{V}$, and if electroneutrality is given by $[\mathrm{e}] = 2[\mathrm{V}]$, we obtain

$$D^\delta_\mathrm{O} = 3D_\mathrm{V}. \tag{31}$$

That is, D^δ_O is given by the diffusivity of the slowest moving defect (in this case: vacancies) multiplied by an 'acceleration factor' (=3) because of the coupling through the Nernst field with the faster moving electrons. In this special case, the activation enthalpy of chemical diffusion is the activation enthalpy of vacancy migration: $\Delta H_{D^\delta_\mathrm{O}} = \Delta H_{\mathrm{mig,V_O}}$
With knowledge of D^δ_O, one can predict the time τ necessary for a sample's nonstoichiometry to proceed, say, to 99% completion. A useful order-of-magnitude approximation for a slab sample of thickness $2l$ is $\tau_{D^\delta} \sim l^2/2D^\delta_\mathrm{O}$. In certain cases, for instance for thin samples, the kinetics of the stoichiometry change are governed by the surface reaction (see Section 5), and the time required is $\tau_{k^\delta} \sim 5l/2k^\delta_\mathrm{O}$. For the intermediate regime, where both bulk diffusion and surface kinetics are important, $\tau_{D^\delta} \sim \tau_{k^\delta}$, a good approximation is $\tau \approx \tau_{D^\delta} + \tau_{k^\delta}$.

4.2 Tracer diffusion

A tracer diffusion experiment refers to the use of radioactive or stable isotopes—chemically identical, labelled species i^*—to examine self-diffusion in condensed matter. Since one can distinguish between i^* and i, the motion of the indistinguishable i particles can be followed with the help of the tracers, i^* [13]. Tracer diffusion experiments are performed at constant sample composition. For our example oxide $\mathrm{MO}_{1-\delta}$, this corresponds in the case of the cation tracer to $\mathrm{MO}_{1-\delta} \rightarrow (\mathrm{M}_{1-\alpha}\mathrm{M}^*_\alpha)\mathrm{O}_{1-\delta}$ and in the case of the anion tracer to $\mathrm{MO}_{1-\delta} \rightarrow \mathrm{M}(\mathrm{O}_{1-\alpha}\mathrm{O}^*_\alpha)_{1-\delta}$.

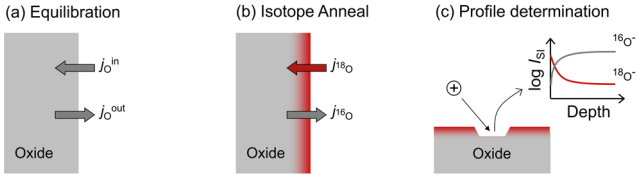


Fig. 6: *The tracer diffusion experiment [14]: (a) An oxide sample is given a pre-anneal in oxygen of normal isotopic abundance at given temperature and oxygen activity in order to equilibrate the sample with the surrounding atmosphere. (b) Subsequently it is annealed, at the same temperature and oxygen activity, in an ^{18}O-enriched gas for a given time. At the sample surface the dynamic equilibrium between gaseous oxygen and oxygen in the sample leads to the incorporation of ^{18}O and the removal of ^{16}O (no net incorporation or removal of oxygen, only the exchange of one isotope for another). Subsequent diffusion of ^{18}O away from the interface and into the solid produces an oxygen isotope profile. (c) The isotope profile in the oxide is commonly determined by an ion-beam-analysis method, such as Secondary Ion Mass Spectrometry (SIMS).*

The only driving force is the gradient in the chemical potential of the tracer (and is thus purely entropic); the chemical potentials of M and O are constant throughout the system.
For many metallic species, convenient radioisotopes are available; radioactive oxygen isotopes, however, are impracticable for diffusion measurements, as their half-lives are at most of the order of minutes. Consequently it is the stable isotope ^{18}O that is used in tracer studies; it can be introduced into a sample either by diffusion annealing in a large volume of ^{18}O-enriched gas at an elevated temperature, or by depositing a thin layer of $M^{18}O$ at room temperature, and then diffusion annealing at an elevated temperature [14]. The former variant is illustrated in Fig. 6.
The measured tracer diffusion coefficient of species i is related to the self-diffusion coefficient through the tracer correlation coefficient f^*,

$$D_i^* = f^* D_i. \tag{32}$$

f^* varies according to the migration mechanism (interstitial, vacancy, colinear/non-colinear interstitialcy) and the geometry of the sublattice on which diffusion takes place (see Table 1). f^* reflects the fact that tracer species do not necessarily perform an uncorrelated random walk. Let us consider the case of vacancy migration on a simple cubic sublattice. The vacancies may move in all six migration directions with equal probability. This is not true, however, for the tracer species: If a vacancy and a tracer have just exchanged places, the most likely jump of the tracer is back to its original position. Its mean square displacement $\langle (R_{i^*})^2 \rangle$ is hence less than that of a random walker, $\langle R_i^2 \rangle$; the tracer correlation coefficient is the ratio of the two, $f^* = \langle (R_{i^*})^2 \rangle / \langle R_i^2 \rangle$, and thus takes values between zero and unity. The detailed analysis with linear irreversible thermodynamics starts from Eq. (24); considers three fluxes, *e.g.* for a vacancy mechanism j_i, j_{i^*} and j_V; and yields f^* in terms of L_{ii} and L_{ii^*} [9, 10]. f^* deviates from unity if L_{ii^*} deviates from zero.
The activation enthalpy of tracer diffusion, $\Delta H_{D_i^*}$ is easily determined by performing measurements as a function of temperature (at constant oxygen partial pressure), but it is not so easy to interpret. For the case of vacancy transport, we find upon combining Eqs. (19) and (32),

$$D_i^* = f^* D_V n_V. \tag{33}$$

Lattice	Mechanism	f^*
diamond	Vacancy	0.5
simple cubic	Vacancy	0.6531
bcc cubic	Vacancy	0.7272
fcc cubic	Vacancy	0.7815
O in ABO_3 perovskite	Vacancy	0.69
any lattice	Interstitial	1
diamond	Interstitialcy (colinear)	0.727

Table 1: *Tracer correlation coefficients for diffusion on various lattices by various mechanisms.*

Since the dependence on temperature of D_{V} and n_{V} can be expressed in exponential functions,

$$D_i^* = f^* D_{\mathrm{V},0} \exp\left(-\frac{\Delta H_{\mathrm{mig,V}}}{k_{\mathrm{B}}T}\right) n_{\mathrm{V},0} \exp\left(-\frac{\Delta H_{\mathrm{gen,V}}}{k_{\mathrm{B}}T}\right), \tag{34}$$

with $\Delta H_{\mathrm{mig,V}}$ being the activation enthalpy of vacancy migration and $\Delta H_{\mathrm{gen,V}}$, reflecting the change in vacancy concentration with temperature, we find

$$\Delta H_{D_i^*} = \Delta H_{\mathrm{mig,V}} + \Delta H_{\mathrm{gen,V}}. \tag{35}$$

Consequently, interpretation of the measured activation enthalpy of tracer diffusion requires quantitative knowledge of the defect chemistry (how exactly does the relevant defect concentration vary with temperature?). $\Delta H_{\mathrm{gen,V}}$, it should be noted, can be positive ($[\mathrm{V}]$ increasing with increasing temperature), negative ($[\mathrm{V}]$ decreasing with increasing temperature) or zero ($[\mathrm{V}]$ independent of temperature); it can take values up to several eV, and hence, depending on the particular case, it may be much larger than, much smaller than or even comparable to $\Delta H_{\mathrm{mig,V}}$.

4.3 Conductivity

In a sense, determining the electrical conductivity of a sample is the simplest of the three transport experiments. One applies an electrical field $E = -\nabla\phi$ to a sample and measures the resulting electrical current density I. The electrical conductivity σ is obtained from Ohm's law

$$I = -\sigma\nabla\phi. \tag{36}$$

There are, however, several possible complications to this simple experiment. First, and most important, is that all mobile charged species contribute to the measured conductivity, all ionic and all electronic charge carriers, each conductivity contribution being the product of the concentration $[i]$, charge $z_i e$ and mobility u_i

$$\sigma_{\mathrm{tot}} = \sum_i \sigma_i = \sum_i [i] z_i e u_i. \tag{37}$$

Generally, the mobilities of electronic charge carriers are orders of magnitude larger than those of ionic charge carriers; hence a small concentration of electrons or holes (minority defects) may provide the dominant contribution to the electrical conductivity. To isolate the ionic contribution, one may have to use the appropriate electron-blocking but ion-conducting electrodes, or

else determine, in addition to the total conductivity, the ionic transference number $t_{\rm ion}$, that is, the proportion of the conductivity carried by the ions, $t_{\rm ion} = \sigma_{\rm ion}/\sigma_{\rm tot}$. The second complication is that the sample's electrical response may be dominated by that of the electrodes, for instance, and obtaining the bulk contribution requires either the use of 4-point measurement geometries in dc mode or frequency-dependent impedance spectroscopy studies. Third, it is only for small driving forces (see below) that Eq. (36) constitutes a linear law, that is, the current density I is proportional to the driving force $-\nabla\phi$, with the constant of proportionality, the measured conductivity, being independent of driving force.
Having determined the ionic conductivity of the bulk phase in the linear regime, one can now calculate the conductivity diffusion coefficient with the aid of the Nernst–Einstein equation,

$$D_i^\sigma = \sigma_i \frac{k_{\rm B}T}{[i](z_i e)^2}. \tag{38}$$

There are various routes to derive Eq. (38), but all of them, it is emphasised, assume the charge carriers under consideration to be non-interacting and dilute [5, 11]. From linear irreversible thermodynamics, for example, one considers a sample of uniform composition ($\nabla\mu_i = 0$) and temperature ($\nabla T = 0$), to which a gradient in the electrical potential $\nabla\phi$ (as the sole thermodynamic driving force) is applied. Comparison of Eq. (27), which is only valid for dilute, non-interacting charge carriers, with Eq. (36) yields Eq. (38). It is to be noted that, although there are various diffusion coefficients, with $D_{\rm ion} \neq D_{\rm def}$ in general, there is only one measured conductivity, *i.e.* $\sigma_{\rm ion} \equiv \sigma_{\rm def}$. Thus, from the measured conductivity, one can calculate $D_{\rm ion}$ if one knows $[{\rm ion}]$, and $D_{\rm def}$ if one knows $[{\rm def}]$. It is mentioned that the ratio of D_i^σ and D_i^* (two diffusion coefficients that can be independently measured) is called the Haven ratio [11]

$$\frac{D_i^*}{D_i^\sigma} = H_R, \tag{39}$$

and that for dilute, non-interacting defects, $H_R = f^*$.
Lastly, we return to the topic of ion migration under high fields. Let us consider the migration of a positively charged interstitial ion in one dimension. In the absence of an applied field, Fig. 7(a), the ion will jump to vacant sites in the forward to backwards directions with equal probability, $\nu \exp\left[-\Delta G_{\rm mig}/k_{\rm B}T\right]$. The field alters these probabilities by altering the barriers, as indicated in Fig. 7(b): it increases the probability of motion in the direction of the field to $\nu \exp\left[-(\Delta G_{\rm mig} - \frac{1}{2}aez_iE)/k_{\rm B}T\right]$ and decreases the probability of motion in the opposite direction in an analagous fashion. The net motion in the direction of the field is proportional to the difference in forward and backward rates. The detailed treatment yields the current density I [23, 24]

$$I = z_i e[i] a\nu \exp\left(-\frac{\Delta G_{\rm mig}}{k_{\rm B}T}\right) 2\sinh\left(\frac{z_i eaE}{2k_{\rm B}T}\right). \tag{40}$$

For small fields, specifically for $|az_ieE| \ll 2k_{\rm B}T$, Eq. (40) reduces to the linear law of Eq. (36) with a field-independent conductivity. For large fields, on the other hand, the jumps of increased probability dominate (*e.g.* for positive ions in the direction of the field), and the current density then displays an exponential variation with the field. In other words, the measured conductivity depends strongly on the field, $I/E = \sigma(E)$, and the field-induced enhancement of the conductivity over the low-field value can be orders of magnitude. The fields required to see significant effects, however, are large: to increase the oxygen-ion conductivity of an oxide (with a typical oxygen-ion jump distance of $a = 0.3$ nm) at room temperature by 25%, one

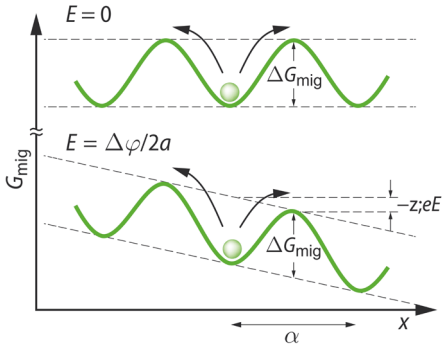


Fig. 7: *Schematic illustration of a positive interstitial ion overcoming an activation barrier of migration ΔG_{mig}: (a) Gibbs energy profile at zero applied field. (b) Gibbs energy profile at non-zero applied field.*

requires a field of $E \approx 1$ MV cm^{-1}. (And a field of $E \approx 2$ MV cm^{-1} will yield an increase in conductivity of 220%.) Such fields are close to the fields at which dielectric breakdown occurs in macroscopic samples; at nanoscale distances, though, this limit may be shifted to even higher values. Nevertheless, it is worth noting that ionic transport can be accelerated exponentially by temperature and/or electric field.

5 Mass transport along and across extended defects

Real oxide samples, one should recognise, are not single crystals containing only point defects: extended defects, such as dislocations and grain boundaries, will in general also be present. In addition, real samples are finite in extent, and thus are bounded by surfaces or interfaces with other phases. A variety of paths may therefore be available for diffusing species (see Fig. 8);

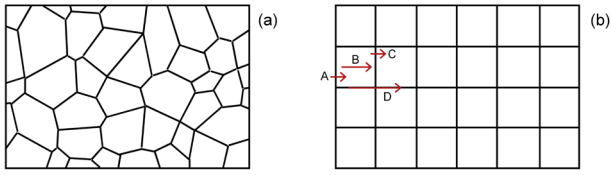


Fig. 8: *Mass transport processes in a polycrystal. (a) Cross-section through a polycrystalline solid. (b) The brick-layer model, an idealised representation of the microstructure shown in (a). The arrows indicate possible transport processes, with the length of an arrow being inversely proportional to the resistance of the associated process. A – hindered transport across a surface; B – transport in the grain bulk; C – hindered transport across a grain boundary; D – enhanced transport along a grain boundary.*

one differentiates between: 1) bulk diffusion (also termed volume or lattice diffusion), which refers to mass transport within a single grain; 2) grain-boundary diffusion, which refers to mass transport along the region of crystallographic misorientation between two grains; 3) dislocation or 'pipe' diffusion, which refers to mass transport along dislocations; and 4) surface diffusion, which refers to mass transport along a crystal surface.

For a given material, the diffusion coefficient of component i along a grain boundary, dislocation or surface (D_i^{gb}, D_i^{dis}, D_i^{s} respectively) will vary according to the structural characteristics of the extended defect. In the case of planar defects (grain boundaries and surfaces), the diffusion coefficient will be a function of the interface (mis)orientation; D_i^{gb}, for example, will vary with the parameters that characterise the grain boundary: the tilt and twist axes, the tilt and twist angles, and the interface plane. Measurements on polycrystals will thus provide an average over all the grain boundaries contained within the investigated volume. In the case of line defects, D_i^{dis} will vary according to the dislocations' character (edge/screw) and its Burgers vector. These alternative paths offered by extended defects become important, if diffusion in the bulk is slow. In such cases, this fast-path diffusion, or short-circuit diffusion, may contribute significantly to mass transport or may even govern the overall behaviour.

Why should diffusion along these extended defects occur faster than in the bulk? The atomic arrangements within grain boundaries and within dislocation cores is considered to be more open than in the bulk phase, suggesting less hindrance for the migrating species (see Fig. 4) and thus accelerated rates of mass transport. Although there is much experimental data that confirms this picture for metallic systems [25–27], it is far from certain that this picture is also universally applicable to oxides. In an (ionic) oxide, an ion diffusing along an extended defect may have to pass, in a migration jump, ions of the same polarity—a process that is avoided in the bulk phase. Furthermore, the intrinsic structure of the extended defects may offer preferential sites for point defects: as a result there may be a high concentration of defects within the extended defects, but they are locked into the structure and thus essentially immobile [28]. A more open arrangement within the extended defect does not, therefore, guarantee accelerated rates of ion transport. In addition, the extended defects may be electrostatically charged, with global charge neutrality being satisfied by attendant, enveloping tubes of space charge in which the concentrations of mobile, charged point-defects are modified drastically from their values in the electroneutral bulk [29].

In addition to fast-path diffusion processes that take place in parallel to diffusion in the bulk

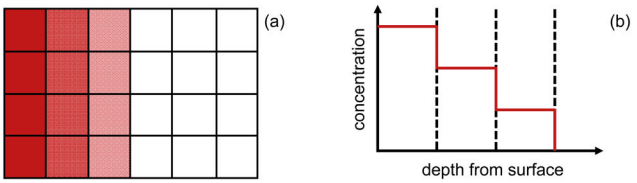


Fig. 9: *Diffusion in a polycrystalline sample with grain size w in which transport across grain boundaries is slow ($k_i^{\mathrm{gb}} \ll D_i/w$) and diffusion along grain boundaries is negligible. (a) Cross-section of a polycrystal: diffusant enters the sample from the left. (b) Concentration profiles across the polycrystal, showing the drops in concentration (Δc_i) at the grain boundaries (whose positions are indicated by the dashed, vertical lines).*

phase, there may be slower processes in series with bulk diffusion (see Fig. 8). Hindered mass transport across interfaces (surfaces, grain boundaries) constitute such slower serial processes, and they may also influence (or even govern) the overall diffusion behaviour. Here, because the concentration is not a continuous function across the interface, one cannot consider a gradient in concentration, and thus one cannot define a diffusion coefficient, as in Eq. (1). Instead, one describes the flux across the interface in terms of a transfer coefficient k_i and the drop in concentration across the interface, Δc_i, in the direction of the flux:

$$j_i = k_i \Delta c_i. \tag{41}$$

The case of transport being limited by grain boundaries is illustrated schematically in Fig. 9.
Why may transport across an interface be hindered? What does k_i refer to, on a microscopic level? There are several possible causes for k_i taking a finite value. At a grain boundary, for example, the crystallographic mismatch between the two grains may conceivably result in the matter flux being diminished, either because of the considerable perturbations of the bulk structure at the interface itself or because of differences in the orientations of the grains in layered structures. The magnitude of such effects are probably small, though. In contrast, huge effects are observed, where space-charge layers, depleted of mobile charge carriers, are present at the grain boundaries [8]. Huge effects are also observed in polycrystalline samples, in which a second phase covers each grain; such second phases are often SiO_2-based compounds that exhibit much lower rates of oxygen transport [30].
In the case of oxygen transport across a surface, that is, across a gas|solid interface, the transfer coefficient $k_{\mathrm{O}}^{\mathrm{s}}$ characterises the exchange flux of the dynamic equilibrium between oxygen in the gas phase and oxygen in the solid. The forward reaction, for example, requires, in addition to several charge transfer steps, the adsorption and the dissociation of oxygen molecules on the surface, and the incorporation of the resulting oxygen moiety into the crystal lattice. The most likely rate determining step is either dissociation or charge transfer leading to dissociation [31].
Thus, in a polycrystal there is a network of serial and parallel mass-transport processes that may be operative. In the following sections I present mathematical models for describing (a) fast-path diffusion along extended defects, such as grain boundaries and dislocations and (b) hindered mass transport across grain boundaries and surfaces.

5.1 Accelerated transport *along* extended defects

The standard model for describing fast grain-boundary diffusion in a polycrystalline sample treats the system as thin grain-boundary slabs of width w^{gb} in which the diffusion coefficient D_i^{gb} is greater than the diffusion coefficient D_i in the bulk grains of width w [33]. The overall behaviour that one observes depends on a number of parameters, such as the diffusion coefficients D_i^{gb} and D_i, the diffusion time and the grain size. Equivalently, the case of fast diffusion along dislocations is treated in terms of tubes of diameter r^{dis}, present at a dislocation density of d, in which diffusion occurs at an enhanced rate D_i^{dis}. Based on the distance over which the bulk structure is perturbed, w^{gb} and r^{dis} are generally assumed to be *ca.* 1 nm. If, however, space-charge zones in which the relevant defects are accumulated are present at such extended defects, the effective values for w^{gb} and r^{dis} will be much larger. Following the example of Atkinson [12], I discuss the three characteristic cases identified by Harrison [32] (see Fig. 10) in order of increasing diffusion time, that is, in reverse order.
Type C kinetics are observed for short diffusion times, for which the diffusant has had insufficient time to penetrate a significant distance into the bulk phase. Diffusion, therefore, takes

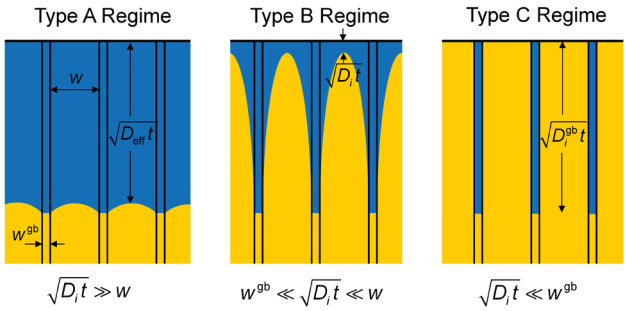


Fig. 10: *Illustration of three regimes of diffusion kinetics in a polycrystal identified by Harrison [32]. The polycrystal consists of grains of width w and diffusion coefficient D_i and grain boundaries of width w^{gb} and diffusion coefficient D_i^{gb}. The three regimes are defined by the inequalities shown in the bottom line of the figure (t is the diffusion time). Adapted from [12].*

place solely along the grain boundaries, without any flux leakage into adjacent grains. The concentration profile that is obtained, for a constant diffusion source, say, follows a complementary error function [Eq. (8)], with the measured diffusion coefficient corresponding to D_i^{gb}.

At longer times, diffusion in the lattice cannot be ignored and, provided the boundaries act independently, type B diffusion kinetics are observed. 'Independently' means that diffusant that comes down one short-circuit path and enters the grains is unlikely to reach another short-circuit path. In concentration profiles, fast short-circuit diffusion makes itself apparent as a 'tail', following the usual bulk diffusion profile. For independent grain boundaries, the tail is linear in a plot of $\ln c$ *vs.* $x^{6/5}$ [34], and analysis of the tail's slope yields the product $D_i^{gb} w^{gb}$;

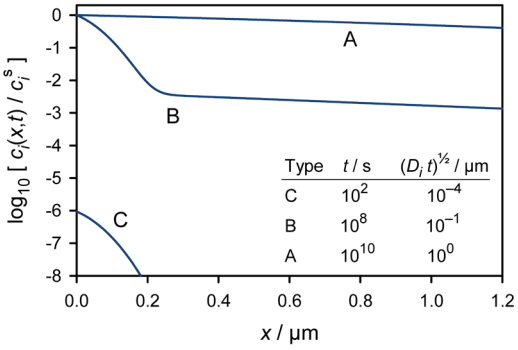


Fig. 11: *Simulated concentration profiles obtained at various times for a single crystal with dislocation density $d = 3 \times 10^7$ cm^{-2}, dislocation radius $r^{dis} = 1$ nm, bulk diffusion coefficient $D_i = 10^{-18}$ cm^2 s^{-1}, and dislocation diffusion coefficient $D_i^{dis} = 10^{-12}$ cm^2 s^{-1}. Adapted from [12].*

for independent dislocations, the tail is linear in a plot of $\ln c$ *vs.* x [35], and analysis of the tail's slope yields the product $D_i^{\mathrm{dis}} r^{\mathrm{dis}}$.
At even longer times, the diffusion fringes around neighbouring grain boundaries overlap extensively and a diffusing moiety may visit many grains and grain boundaries during the diffusion time. The grain boundaries are not acting independently. This is type A diffusion and the concentration profile follows, for a constant diffusion source, a complementary error function; that is, it is indistinguishable from pure bulk diffusion. In this case, however, the effective diffusion coefficient D_{eff} is a function of D_i and D_i^{gb}.
The whole picture becomes far more complicated if space-charge zones that are depleted of the mobile defects are present. There may be fast diffusion along the interface itself, but the flux is prevented from leaving the interface by the depletion space-charge
zones. In this case the diffusion kinetics will not correspond to Harrison type A, B or C.

5.2 Hindered transport *across* extended defects

Owing to its importance for ionic oxides, we consider specifically in this section transport across interfaces being hindered by depletion space-charge layers. This can be investigated experimentally by means of all three transport experiments (chemical, tracer, conductivity), but the tracer approach currently offers one major advantage: it is capable of resolving the profile within the space-charge zone. This is shown in Fig. 12 for the case of a depletion space-charge zone at a surface.

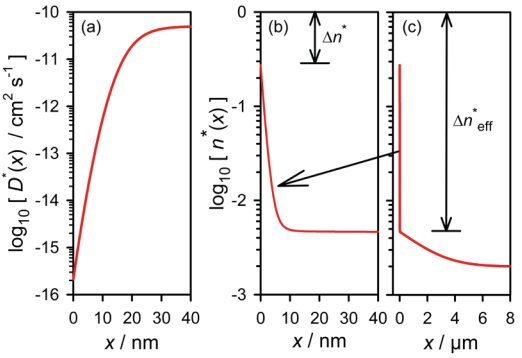


Fig. 12: *Isotope transport through an equilibrium surface space-charge layer depleted of oxygen vacancies [36, 37]. (a) Local variation of the oxygen tracer diffusion coefficient,* $D^*(x) = f^* D_{\mathrm{V}} n_{\mathrm{V}}(x) \approx D^*(\infty) \exp[-2e\phi(x)/k_{\mathrm{B}}T]$*, that arises from oxygen-vacancy depletion near the surface. Solving Eq. (3) with this spatially variant* $D^*(x)$ *yields the isotope profile shown in (b) [the first 40 nm], and in (c) [the entire profile].*

Observed with moderate depth resolution [Fig. 12(c)], the profile shows a drop in isotope fraction, $\Delta n^*_{\mathrm{eff}} = j^*/k_{\mathrm{eff}}$, that is the combined effect of the limited surface-reaction kinetics and the depletion space-charge layer. At ultra-high depth resolution [Fig. 12(b)], the profile within the space-charge layer becomes apparent, as well as the drop in isotope fraction that arises from the surface-reaction kinetics, $\Delta n^* = j^*/k^{\mathrm{s}}$.

References

[1] R. A. De Souza, *Adv. Funct. Mat.*, 2015, **25**, 6326–6342.

[2] R. A. De Souza, in *Resistive Switching: From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications*, ed. D. Ielmini and R. Waser, Wiley-VCH, Weinheim, 2016, pp. 125–164.

[3] P. Shewmon, *Diffusion in Solids*, TMS, Pennsylvania, 1989.

[4] H. Mehrer, *Diffusion in Solids*, Springer, Berlin Heidelberg New York, 2007.

[5] H. Schmalzried, *Festkoerperreaktion*, Verlag Chemie, Weinheim, 1971.

[6] R. W. Balluffi, S. M. Allen and W. C. Carter, *Kinetics of Materials*, John Wiley & Sons, New Jersey, 2005.

[7] R. J. Borg and G. J. Dienes, *An Introduction to Solid State Diffusion*, Academic Press, San Diego, 1988.

[8] J. Maier, *Physical Chemistry of Ionic Materials: Ions and Electrons in Solids*, J. Wiley & Sons, Chichester, 2004.

[9] R. E. Howard and A. B. Lidiard, *Rep. Prog. Phys.*, 1964, **27**, 161.

[10] C. Wagner, *Prog. Solid State Chem.*, 1975, **10, Part 1**, 3 – 16.

[11] G. E. Murch, in *Phase Transformations in Materials*, ed. G. Kostorz, Wiley-VCH, Weinheim, 2001, pp. 173–238.

[12] A. Atkinson, *Solid State Ionics*, 1984, **12**, 309–320.

[13] M. Martin, in *Diffusion in Condensed Matter*, ed. P. Heitjans and J. Kaerger, Springer, Berlin Heidelberg New York, 2005, pp. 211–249.

[14] R. A. De Souza and M. Martin, *MRS Bulletin*, 2009, **34**, 907–914.

[15] J. Janek, J. Sann, B. Mogwitz, M. Rohnke and M. Kleine-Boymann, *J. Korean Ceram. Soc.*, 2012, **49**, 56–65.

[16] T. Lee, H.-S. Kim and H.-I. Yoo, *Solid State Ionics*, 2014, **262**, 2–8.

[17] J. Crank, *The Mathematics of Diffusion, 2nd edition*, Oxford University Press, Oxford, 1975.

[18] G. H. Vineyard, *J. Phys. Chem. Solids*, 1957, **3**, 121–127.

[19] S. R. de Groot and P. Mazur, *Non-Equilibrium Thermodynamics*, North Holland, Amsterdam, 1962.

[20] D.-K. Lee and H.-I. Yoo, *Phys. Rev. Lett.*, 2006, **97**, 255901.

[21] C. Wagner, *Prog. Solid State Chem.*, 1972, **7**, 1–37.

[22] H.-S. Kim and H.-I. Yoo, *Phys. Chem. Chem. Phys.*, 2011, **13**, 4651–4658.

[23] E. Verwey, *Physica*, 1935, **2**, 1059–1063.

[24] N. F. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystal, 2nd edition*, Oxford University Press, Oxford, 1950.

[25] I. Kaur, Y. Mishin and W. Gust, *Fundamentals of Grain and Interphase Boundary Diffusion*, J. Wiley & Sons, Chichester, 1995.

[26] N. Peterson, *Int. Mater. Rev.*, 1983, **28**, 65–91.

[27] R. Balluffi and R. Mehl, *Metall. Trans. A*, 1982, **13**, 2069–2095.

[28] R. A. De Souza, *Phys. Chem. Chem. Phys.*, 2009, **11**, 9939–9969.

[29] V. Metlenko, A. Ramadan, F. Gunkel, H. Du, H. Schraknepper, S. Hoffmann-Eifert, R. Dittmann, R. Waser and R. De Souza, *Nanoscale*, 2014, **6**, 12864–12876.

[30] J.-H. Lee, *Monatsh. Chem.*, 2009, **140**, 1081–1094.

[31] R. A. De Souza, *Phys. Chem. Chem. Phys.*, 2006, **8**, 890–897.

[32] L. G. Harrison, *Trans. Faraday Soc.*, 1961, **57**, 1191–1199.

[33] J. C. Fisher, *J. Appl. Phys.*, 1951, **22**, 74–77.

[34] A. D. Le Claire, *Br. J. Appl. Phys.*, 1963, **14**, 351.

[35] A. D. Le Claire and A. Rabinovitch, *J. Phys. C: Solid State Phys.*, 1981, **14**, 3863.

[36] R. A. De Souza and M. Martin, *Phys. Chem. Chem. Phys.*, 2008, **10**, 2356–2367.

[37] R. A. De Souza, V. Metlenko, D. Park and T. E. Weirich, *Phys. Rev. B*, 2012, **85**, 174109.

A5 Phase Transitions

Chr. Pithan
Peter Grünberg Institute PGI-7
Institute for Electronic Materials
Forschungszentrum Jülich GmbH

Contents

1 Abstract **2**

2 Introduction: a historical retrospect **2**

3 Elementary phase diagrams **3**

4 The state forms of matter & related phase transitions **6**

5 Types of heterogeneous phase transitions **7**
5.1 Phase transitions involving the solid and liquid state 7
5.2 Phase transitions in the solid state 8
5.3 Concluding remarks on the relevance of phase transitions 14

6 Fundamental thermodynamic considerations **15**
6.1 Definitions 15
6.2 The Gibbs phase rule 17
6.3 Order parameters 18
6.4 Thermodynamical functions, process parameters and equilibrium 19

7 Kinetic aspects: nucleation and growth **22**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Abstract

The objective of the present tutorial is to review and summarize the fundamental principles of phase transitions, which represent transformations of one state of matter to another one. A particular focus is placed on solid state science and engineering, as far as needed to provide a sound basis for helping to find to a deeper understanding of the later and further more advanced and specific chapters of this volume. Certainly, the limited possibilities to treat this subject comprehensively are obvious but it is the authors hope and wish that the present chapter will somehow be instructive and also stimulate the reader independently of his or her background of knowledge to find a deeper path to comprehend the field. Covered are, after a brief introduction, reminding a few very basic aspects using simple illustrating examples, real material systems, thermodynamics and kinetics of phase transitions It might appear boring to the advanced reader – but a fresh-up never hurts. It was always tried, wherever possible, to concentrate on solid-solid phase transitions. However, an overview on a multitude of different categories and types of phase transitions and transformations is presented too, even if they cannot be discussed in full detail in the framework of such a restricted examination. In the following, some necessary definitions of the most relevant terms and notions of thermodynamics that are used to describe phase transitions in connection with the energetic background of these phenomena are presented. In this context, also the essentials of the mathematical treatment of thermodynamics needed to comprehend its essential principles are treated. A more specific emphasis was put on a simplified presentation of the phenomenological so-called Landau model. This phenomenological approach appears to be very helpful to describe second order phase transitions such as ferroelectric, magnetic and even other transformations, representing an important branch of physical ordering phenomena, in which quite a multitude of effects turn out to be very significant also in modern information technology. Finally, the current presentation closes with some considerations about the kinetics of phase transitions regarding the nucleation and growth, the initial stages of transformations of matter.

2 Introduction: a historical retrospect

The question about, what our materialistic world really consists of has fascinated and mystified humanity since a long time of its history. One of the very early thinkers, who tried to summarize the first thoughts regarding this problem was the ancient Greek philosopher Aristoteles (384 – 322 b.c.), Fig. 1. In the view of this universal sage scholar, all forms of matter are distinguished by only *"four types of appearance"***.** These were *"Fire"*, *"Air"*, "*Water"* and *"Earth"* a concept that survived for a long time over the centuries of the medieval and modern ages even through – at least partially – the upcoming age of Enlightenment and the period of the advancement of modern natural sciences. It is interesting to note that also in very far distant other Asian cultures such as in China or in Japan the belief about these elementary states of matter evolved. Still even currently some weekdays are named after these aspects in those nations. Tuesday is the day of *"Fire"*, Wednesday the day of *"Water"* and Saturday represents the day of *"Earth"*.

A more scientific treatment of the physical modifications of matter was only possible with the invention and development of the first thermometers and the definitions of different temperature scales. With the ending 17th Century and beginning 18th Century scientists such as Robert

Boyle (1627 – 1691), Edmé Mariotte (1620 – 1684) and Louis Gay-Lussac (1778 – 1850) systematically studied by experiments, how gases change regarding to temperature, pressure and volume. This might have be the birth hour of modern thermodynamics.

Fig. 1: *Aristoteles (384 – 322 b.c.) on left side and the categorization of four types for the appearance of matter ("Fire", "Air", "Water" and "Earth") on the right.*

It is worthwhile to mention that even the library of Sir Isaac Newton (1642 – 1726), probably representing one of the most prominent fathers of modern physical science and a contemporary of the above-mentioned early scholars of thermodynamics, still contained a quite considerable quantity of published scriptures referring to Alchemy. Finally, it is therefore quite remarkable and fascinating that Aristoteles already recognized with his only limited possibilities of methodical scientific observations the fundamentals of what we know nowadays naturally know as the three state forms of matter gas, liquid and solid and of plasma, as the fourth form.

3 Elementary phase diagrams

A very simple consideration arising from observations of nature teaches us since our youngest age – rain or snow falling from clouds in heaven – that water (H_2O) exists in three different optically and tactically distinguishable forms: vapour (water gas; steam is strictly speaking a mixture of water vapour and very tiny droplets of liquid water), liquid water and finally solid ice. Without the existence of these modifications, life on our planet would not be possible. There would be no climate and consequently photosynthesis providing the growth of vegetation and thus the basis of our nourishment could not exist.

Furthermore, the delicate equilibrium between the gases oxygen O_2 and carbon dioxide CO_2 in the earth's atmosphere allowing us to breathe would not subsist.

Whether water in its pure form is liquid, gaseous or solid depends on external conditions defined by pressure p and temperature T and can be visualized by a so-called phase diagram. Such graphical representations show the stability areas of all three phases in dependence of these parameters pressure p and temperature T (Fig. 2), which are found on almost every weather forecast map and essentially determine our climate, whether it is cold or hot or either whether it is influenced by high- or low-pressure regions.

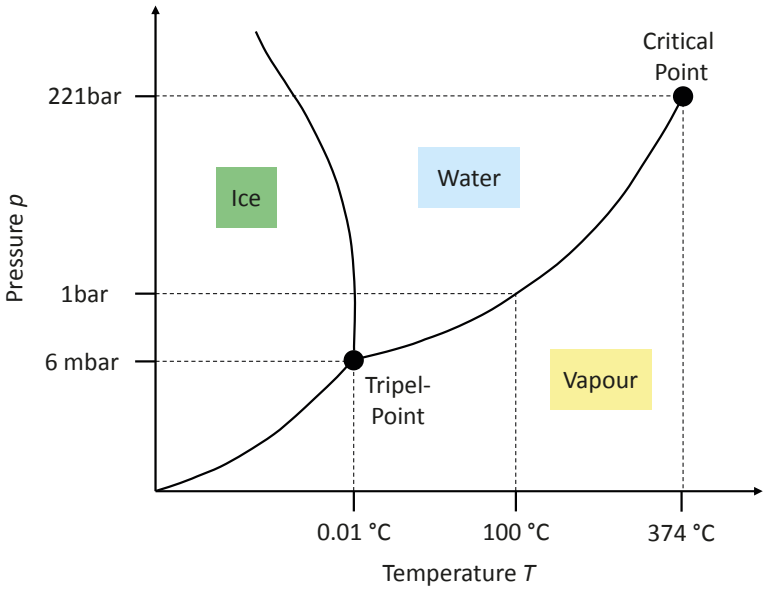


Fig. 2: *The pressure-temperature (p-T) – phase diagram of water H_20.*

The phase diagram shown in Fig. 2 contains three single-phase regions, marking the ranges of existence of liquid, gaseous and solid water. Within each of these three single-phase regions, all mentioned modifications exist independently, meaning that any small variation within a certain range in the parameters temperature T and pressure p will not result in the formation of even a small amount of another state of water. The coexistence of two different forms of water (liquid & solid, liquid & gaseous or solid & gaseous) is only possible at the lines separating the respective regions of stability of the individual one-phase regions. This implies that such so-called two-phase equilibria can only be maintained upon the change of one arbitrarily chosen external parameter, if the other one is correspondingly adopted. One example, for illustration, referring to Fig. 2 again: pure water boils at 100 °C at 1 bar of pressure. If the pressure is reduced, water will start to boil at a lower temperature. Equally, single phase superheated water vapour (above 100 °C) will not condense at a pressure of 1 bar but at a respectively higher value of external pressure. In the case of heated liquid water at a constant ambient pressure of 1 bar applying caloric energy, the initial temperature will increase as long as the respective two phase equilibrium of water and vapour has not yet been reached, meaning up to the moment as soon as the first smallest amount of the gas-phase is released. Any further introduction of heat will leave the system at constant pressure (1 bar) and temperature (100 °C).

Gradually more and more water will evaporate under these fixed conditions upon further heating, until all of the liquid will be transformed in vapour. When all liquid water has vanished and transforms into vapour, further introduction of heat to the system will then lead to an increase of temperature again. The reason for this behaviour is that gradually intermolecular bonds in the liquid phase are broken and that water molecules move more vividly and intensively in the gaseous state: the density is decreased and equally disorder is increased. The differences in the two different energetic states liquid and gaseous requires the introduction of latent heat, and therefore the temperature remains constant (at a fixed given value of pressure, here in this example p = 1 bar) until full vaporization is achieved.

The equilibrium, at which all three forms of water can coexist simultaneously, is defined by the so-called triple point at 6 mbar of pressure and 0.01 °C. At no other combination of pressure and temperature solid, liquid and gaseous pure water are in equilibrium.

At the critical point of water (p = 221 bar and T = 374 °C) vapour and liquid water are not distinguishable anymore: at these conditions the gaseous phase has been compressed so far and the liquid phase has been diluted by heating to that extend that they both appear to be identical: their densities seem to be the same.

The presented essential effects observed for a simple system as pure water – as will be seen later in the framework of this tutorial – illustrate that the state of a certain system cannot be varied arbitrarily but that it depends on certain thermodynamical constraints, such as whether the system under consideration is chemically pure or whether it consists of several chemical elements or compounds and how many phases are to coexist.

A second, schematic representation – of a unary, pure system, such as water, shown in Fig. 3 – illustrates the invariance of certain state variables during a transition from one condition into another one.

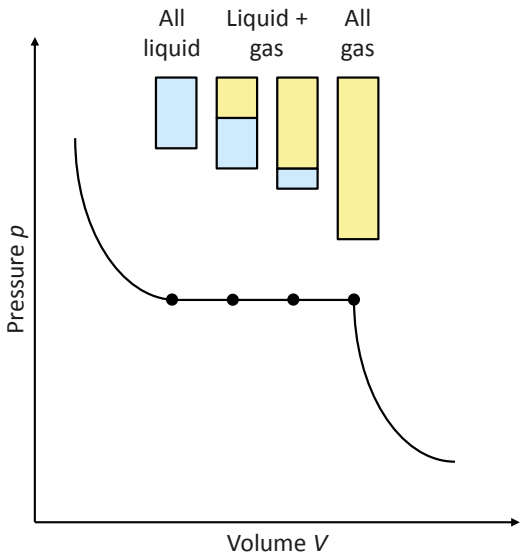


Fig. 3: *The pressure-volume dependency during the liquefaction of a gas.*

In the present case, the liquefaction of a gas by compression is considered: the transformation from a gas into a liquid by compression. At large volumes V and small values of pressure p the gas will be uniform and homogeneous, as indicated by the yellow bar at the right side of the diagram, until a certain critical pressure is reached, at which the first small amounts of water will start to condense. Upon further reduction of the volume more gas will liquefy (symbolised by the blue bar increasing in size), but at the same time pressure will remain constant until all rests of the remaining gaseous phase have disappeared. If the volume is continued to be decreased in the purely liquid state, the pressure will start to rise again.

4 The state forms of matter & related phase transitions

More generally, the state forms of matter for all kind of substances can be summarized with the schematic representation of Fig. 4. They are ordered in this list from the bottom to the top according to decreasing structural order and decreasing density ρ. Condensed forms of matter such as solids or liquids have remarkably higher density than gases. In solids, whether they are crystalline (high translation symmetry) or amorphous (only short range order), ordering of atoms, ions or molecules is relatively high. Upon melting the kinetic energy of such elementary particles increases and the potential interaction- or bonding energy between them is reduced. If a liquid substance evaporates into a gas the average distance between the elementary particles and therefore also their interaction is lessened even more. Further dilution and increase in disorder is achieved if a gas is ionized into plasma.

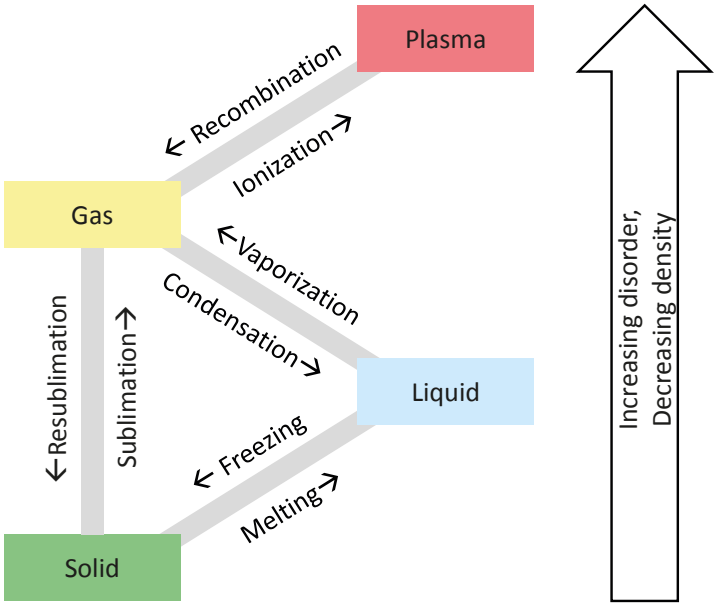


Fig. 4: *Overview on different phase transitions possible between the various state forms of matter, listed from the bottom to the top with increasing disordering and reduced density. The nomenclature for the respective transition processes involved are indicated with arrows next to the corresponding connection lines.*

A solid may be transferred to a liquid by melting. The opposite case is called *solidification* (alternatively *crystallization* or *freezing*), where a melt transfers into a solid. This also includes the situation when a melt is transformed, for instance by rapid quenching into a non crystalline but amorphous state, in which the lack of long-range order of the high temperature liquid phase may be (at least partially) frozen in.

Solids can be directly evaporated into the gaseous state. This kind of phase transition is referred to as *sublimation*. The reverse process by which a gas directly deposits as a solid phase is denoted as *resublimation*, a mechanism particularly relevant for the deposition of thin films (MOCVD, ALD). In the case of thin film deposition by the application of some physical methods, such as magnetron sputtering even deposition from plasma directly into the solid phase can take place.

The transformation of a liquid into a gas is called *vaporization* and the invers process of liquefaction of a gas, *condensation*. If enough energy can be transmitted to a gas, individual molecules can be ionized and form a plasma (*Ionization*). Contrarily free ions can also recombine to gaseous molecules again (*Recombination*).

5 Types of heterogeneous phase transitions

5.1 Phase transitions involving the solid and liquid state

Beyond the elemental transitions between different forms of state of matter some more examples of phase transformations that are exclusively observed in systems consisting of at least more than two chemical elements should not be left unmentioned, because they are of essential importance to materials science and technology. They refer to equilibria between the solid and the liquid state of matter. For the sake of simplicity we here confine our consideration to a two component system (binary system). However, in principle the same treatment can be done for higher ordered systems (ternary, quaternary, quinary ...):

- *Monotectic phase reactions* refer the decomposition of a liquid phase l_1 into a solid α and another liquid phase l_2. Formally, such a decomposition reaction can be expressed as:

$$l_1 \rightleftarrows \alpha + l_2 \tag{1}$$

- *Eutectic phase reactions* describe the decomposition of a liquid phase l into two or more solid phases (α, β) in systems that show limited solubility of their constituent chemical components. The corresponding equation then reads as follows:

$$l \rightleftarrows \alpha + \beta \tag{2}$$

- *Peritectic phase reactions* involve the reaction of a liquid phase l with a solid phase α to a second solid phase β. The reaction formalism is then expressed by the following equation:

$$l + \alpha \rightleftarrows \beta \tag{3}$$

A none totally really extensive list summarizing examples of the most prominent possible phase transitions and transformations in the solid state is given below together with illustrating reaction equations whenever available and illustrative examples of specific material systems are known.

5.2 Phase transitions in the solid state

- *Crystallographic or polymorphic transitions*: these are quite often observed in nature. The most prominent example, probably, because of its historical and technological importance as the basic constituent in the metallurgy of steels is the polymorphism of iron Fe, which exists in different allotropic modifications of it crystal lattice (body centered cubic or face centered cubic structure). A second example is carbon C, which naturally either exists as extremely hard, transparent crystals, with a tetrahedral atomic structure (diamond) or in the more stable and abundant form of graphite, where C atoms are arranged in flat hexagonal lattices (graphene).
- *Spinoidal decompositions* represent the particular case of spontaneous un-mixing of initially compositionally homogeneous but supersaturated solid solutions into two separated different coexisting solid phases within a so-called miscibility gap. The process is solely determined by precipitation through (negative) diffusion, since almost no energetic barrier exists, requiring the formation of stable nuclei of the two new phases. Important examples are many precipitation hardened Al-based alloys, such as for instance the systems Al-Mg, Al-Zn or Al-Li-Sc. Fig. 5 shows an example of an Al_3(Li, Sc) core-shell particle, which precipitated in a highly monodisperse size distribution in Al-Li-Sc alloys. Such precipitations, as they are also found in the classical system Al-Cu (Duraluminum, Alfred Wilm, 1869 – 1937), typically develop in a certain sequence with different states of coherency to the surrounding matrix. Usually initially coherent precipitates are formed, since they possess the smallest lattice mismatch to the original phase. From these, semi-coherent precipitations evolve along certain lattice planes and finally incoherent precipitates, which are essential for increasing the mechanical strength of such alloys by impeding plastic deformation through the motion of dislocations, develop.

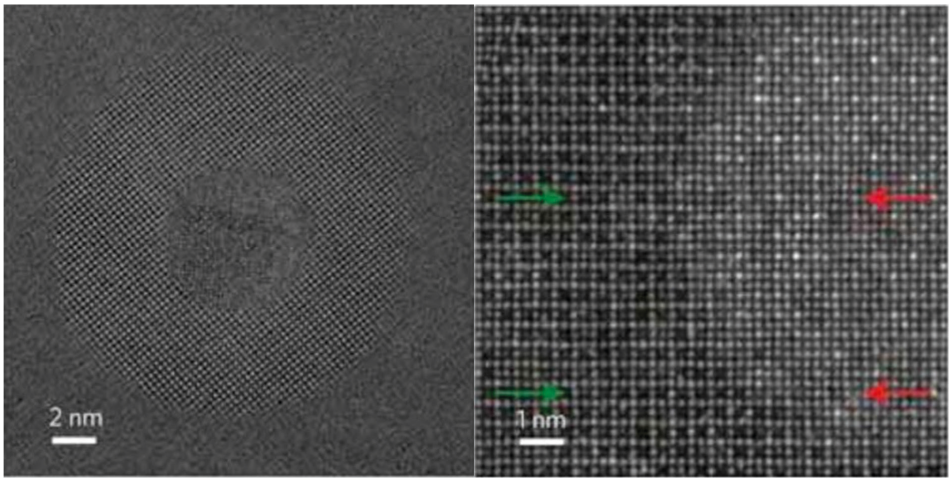


Fig. 5: *High resolution Transmission Electron Microscopic micrograph (left) and HAADF (right) image of a Al_3(Li,Sc) core-shell precipitate and its internal and external interfaces in a Al-matrix [1].*

- *Eutectoid decompositions* are phase separations where one high-temperature solid solution decomposes into two different low-temperature phases through a coupled mechanism of nucleation and growth during cooling. The corresponding reaction equation can be written as:

$$\gamma \rightarrow \alpha+\beta \tag{4}$$

A well-known technically highly relevant example for this decomposition mechanism is the formation of pearlite, a phase mixture of body centered cubic ferrite (α-Fe) and cementite Fe_3C from an initially homogeneous face centered cubic solid solution of austenite (γ-Fe), in which C is completely dissolved in the crystal lattice. Typically, the microstructure of the resulting decomposition product has a lamellar arrangement of both solid phases formed, as the example in Fig. 6 illustrates. Eutectoid reactions play an eminent role in the metallurgy of steels. In the pure Fe-C – system (technical steels usually contain a multitude of additional alloying elements) the eutectoid decomposition of austenite occurs isothermally at 723 °C. Since C has to diffuse to form Fe_3C (25 at.-% C) leaving α-Fe, with almost no C dissolved back, and because critical nuclei have to be formed, this reaction is invariant to temperature consuming latent heat from the initial high temperature phase austenite.

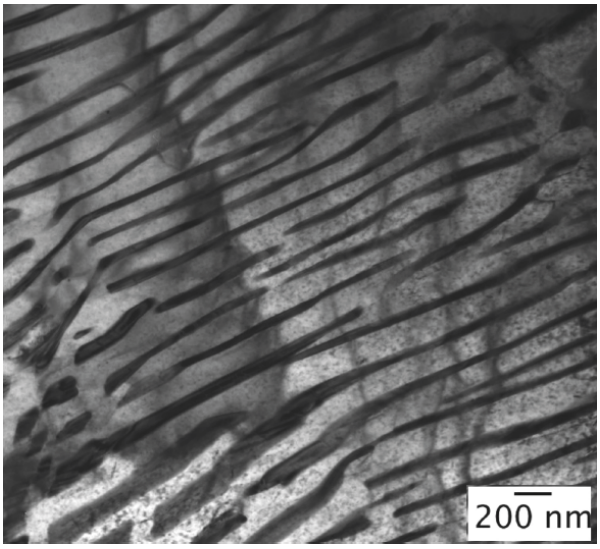


Fig. 6: *Transmission Electron Microscopic image showing the alternate lamellar arrangement of ferrite (α-Fe, light phase) and cementite (Fe_3C, dark phase) in extremely fine arrangement of pearlite with interlamellar spacings of less than 100 nm in a Fe-0.8C-1.6Si-1.9Mn-1.3Cr-0.30Mo steel (concentrations in wt.-%) [2].*

- *Martensitic transformations* are diffusion-less solid-state transformations, in which phase changes occur without long-range diffusion of atoms. Such structural transformations rather involve a cooperative and coordinated type of homogeneous displacement of many atoms. The lattice distortive shear displacements are generally remechanical

lated to the introduction of high amounts of elastic strains, internal interfaces and planar defects such as twins or stacking faults. Typical examples for martensitic transformations are steels, annealed at elevated temperatures in order to stabilize the face centered cubic modification austenite (γ-Fe) and subsequently quenched at such high cooling rate that the stable eutectoid reaction

$$\gamma\text{-Fe} \rightarrow \alpha\text{-Fe} + Fe_3C \tag{5}$$

is suppressed. This phase separation requiring the diffusion of C-atoms and the nucleation of the lamellar Fe_3C-phase cannot occur and instead the lattice of the high temperature austenitic phase transforms via mechanical shearing into a metastable highly distorted microstructure. This process implies a considerable increase in mechanical strengthening (hardening) through the high amount of stored elastic deformation energy and a large quantity of planar lattice defects and interfaces, as shown exemplarily in Fig. 7, which obstruct extensive plastic deformation via dislocational motion. Other technical application involving martensitic transformations are memory shape alloys, for example based on the system Ni-Ti.

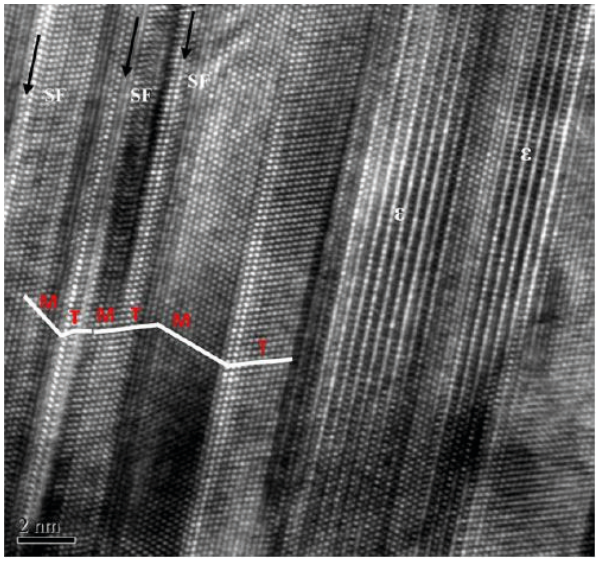


Fig. 7: *High resolution Transmission Electron Microscopic micrograph showing martensite formation in a austenitic steel (Fe-0.027C-0.36Si-1.93Mn-14.35Cr-8.71Ni-0.79Cu, concentrations in at.-%) under a tensile strain of 57 %, containing ultrafine structured features, such as stacking faults (SF) and twins (T)[3].*

- *Peritectoid transitions* represent a rather rare cases of solid-solid phase reactions upon which two high temperature solid phase transform upon cooling to a new solid phase according to the following reaction equation:

$$\alpha+\beta \rightarrow \gamma \tag{6}$$

In terms of heating, this type of equilibrium implies the reaction of two solid phases, which are stable at low temperatures to a new different high temperature solid phase. An example for the resulting microstructure is shown in Fig. 8. representing a Cu-Sn-Al - alloy, where two solid phases α and β_1 react to a new phase β at the interface of both initial high temperature phases.

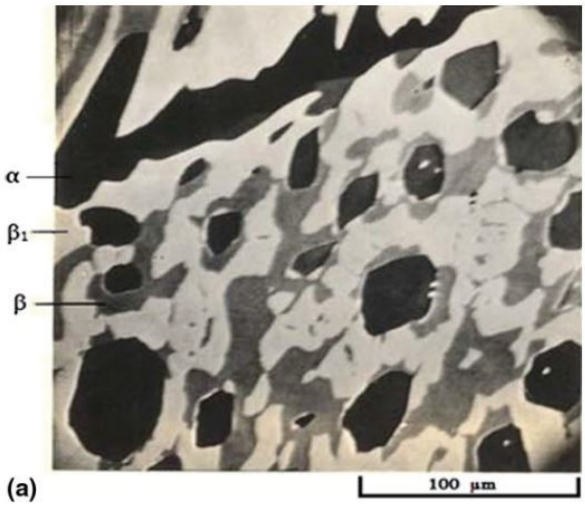


Fig. 8: *Microstructure, of a mixture containing three phases α (dark), β(light) and β_1(dark grey) of a Copper alloy containing 20 at.-% Al and 3 at.-% Sn after prolonged thermal decomposition at 560 %C [4]. The β-phase evolves out of the β_1-phase through a solid-solid state reaction at the surface of grains formed by the α-phase.*

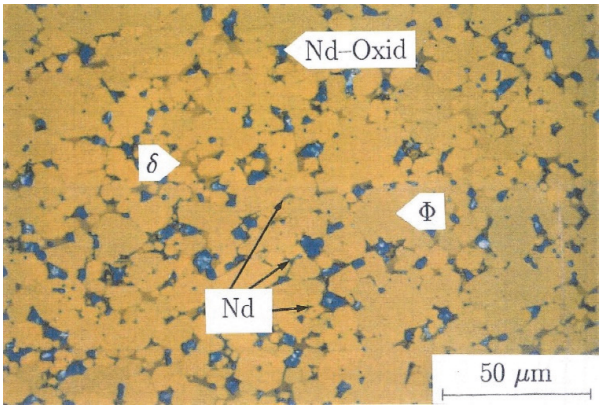


Fig. 9: *Microstructure (light microscopy, bright field) of a sintered permanent magnet with the composition Fe78.1-Nd19.1-B5.4-Ga0.4 (composition in at.-%). During cooling a peritectic reaction of the liquid phase (liquid phase sintering) and the ferromagnetic intermetallic phase Φ ($Fe_{14}Nd_2B$) form another intermetallic phase δ $(Fe,Ga)_6Nd_{14}$ at the surface of the ferromagnetic particles of Φ, improving the magnetic decoupling of these particles and therefor also the coercive field strength [5].*

Technically more relevant however, are the already above mentioned peritectic reactions, where one of the initial phases is liquid. An example shows Fig. 9, chosen from the quaternary system Fe – Nd – B – Ga, which is relevant to permanent magnetic materials.

- *Ferroelectric ordering* involves phase transitions, where a high temperature paraelectric, non-polar phase of higher symmetry transforms upon cooling into a polar polymorph with lower crystallographic symmetry. This transition usually occurs at a certain critical temperature, which is referred to as Curie temperature. In consequence, a lattice distortion of these often solid ionic systems results in the formation of electric dipoles as shown in Fig. 10, representing the tetragonally distorted crystallographic structure of the ferroelectric compound $Pb(Zr,Ti)O_3$.

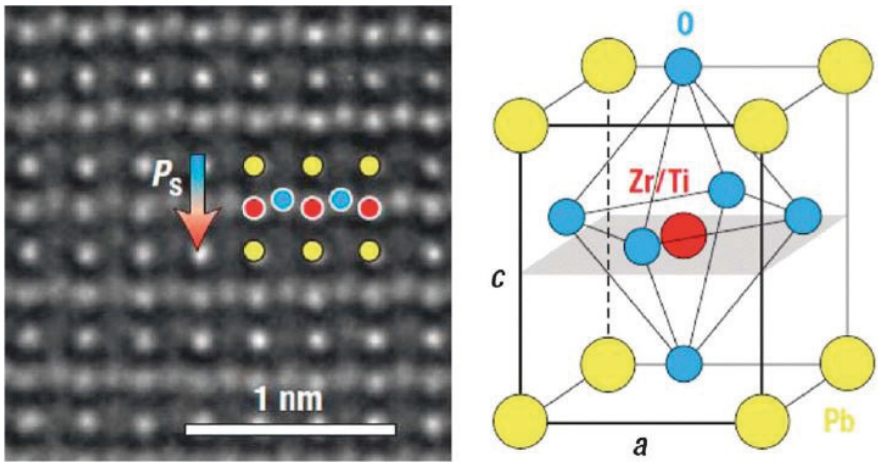


Fig. 10: *Atomic scale TEM image (left) and graphical, schematic representation (right) of the tetragonally distorted crystal structure of ferroelectric $Pb(Zr,Ti)O_3$. [6]. The Pb^{2+}-cations are situated on the corners of the tetragonal unit cell, whereas the O^{2-}-anions are located on its face centers, forming a octahedron that surrounds the central Zr^{4+}- or Ti^{4+}-cation around the middle.*

As can be recognized from Fig. 10 the central cation (in this case Zr^{4+} or Ti^{4+}) is not exactly situated in the center of the tetragonal unit cell in the ferroelectrically ordered state but slightly displaced, giving rise for net dielectric dipole. The displacement can take two possible positions in the present case, which are referred to as the both polarization states of the system. When an electrical field is applied to the material one polarization state can be switched into the other one. Generally, both polarization states are observed in macroscopic systems above a certain critical size. They form so-called domains and can be visualized for instance by Transmission Electron Microscopy, as the example of the ferroelectric compound $BaTiO_3$ shown in Fig. 11 illustrates. Within one individual domain, the direction of polarization is constant, all elementary dielectric moments being ordered in a parallel way, but in neighboring ones, they differ by a certain angle that depends on the system under consideration. If a sufficiently large electric field (in the order of generally a few kilovolts per centimeter) is applied to such a polydomain structure, gradually the differently oriented domains may be turned into the same direction until finally in the saturated state only one single domain exists.

Ferroelectric materials are widely applied in electronic devices such as passive components (capacitors, piezoelectric actuators or sensors and pyroelectric elements.

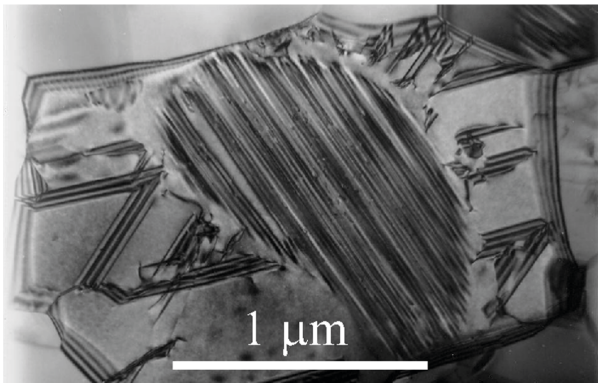


Fig. 11: *Ferroelectric domain structure in a polycrystalline ceramic of the ferroelectric compound $BaTiO_3$ [7].*

The above consideration showed that in ferroelectric materials – at least within some restricted regions of space – individual dielectric dipoles summing up to a macroscopic spontaneous polarization are oriented in a parallel order. In the case of *anti-ferroelectric ordering* this is not the case. Here the orientation relationship is antiparallel from one unit cell to the neighbouring one. An example for an antiferroelectric compound is sodium niobate $NaNbO_3$.

- *Magnetic ordering* relates to ordered structures of atomic magnetic moments, generally in solid materials. Ordering occurs generally from the paramagnetic state into a ferromagnetic, antiferromagnetic or ferrimagnetic phase. Differently to ferro- or antiferroelectrically ordered solids, however, magnetic transitions are not necessarily coupled to a change in crystal symmetry. A classic example is pure iron Fe that forms a room cube centered lattice up to a temperature of 912 °C and which transforms from its low temperature ferromagnetic state to the high temperature ferroelectric state already at a Curie point of 769 °C. Magnetic ordering depends on the interactions of atomic magnetic moments. In the disordered state, the thermally induced vibration of atoms or ions predominates energetically and in consequence, the random orientation distribution of a paramagnetic phase is more stable. If the temperature is reduced also the thermal agitation decreases and at a certain critical temperature (Curie point for ferro- and ferrimagnets and Néel point for antiferromagnets) magnetic interaction prevails. Then a parallel arrangement will establish in the case of ferromagnetic and an antiparallel order for antiferromagnetic interaction. Typical ferromagnetic materials are the transition metal elements Fe, Co, Ni and Mn and many of their alloys, particularly steels, Heusler alloys and the so-called AlNiCo-materials. In addition, many intermetallic phases formed by them and Rare-Earth metals show distinct ferromagnetic behavior.

  Examples for antiferromagnetic are transition metal compounds, especially oxides such as Fe_2O_3 and NiO. They do not reveal a macroscopically measureable spontaneous magnetization because the two oppositely oriented magnetic sublattices compensate each other. Ferrimagnetic materials somehow take an intermediate position in this classifica-

tion. They usually exist as compounds in which different magnetic ions, possessing not the same magnitude of the magnetic moment, also form distinct antiparallel magnetic sublattices.

In contrast to antiferromagnets they therefor sum up to a macroscopic spontaneous magnetization, which however is generally smaller than the ones observed for ferromagnetic materials. Examples are magnetite Fe_3O_4, spinel type ferrites such as $ZnFe_2O_4$ or $MnFe_2O_4$, hexagonal type ferrites such as $BaFe_{12}O_{19}$ or $SrFe_{12}O_{19}$ and garnets, like $Y_3Fe_5O_{12}$.

Ferro- and ferrimagnetic materials also split up into homogeneously ordered magnetic domains, as the illustrative example in Fig. 12 shows.

Fig. 12: *Magnetic domain structure in a cube texture transformer steel [8]. The magnetic field strength is gradually increased from zero field (left) to applied field H (center and right). Cleary the variation of the domain pattern upon magnetization can be recognized.*

5.3 Concluding remarks on the relevance of phase transitions

All different kinds of phase transitions are of eminent importance in solid-state sciences and technologies: they represent the basis of the richness in different engineering structural as well as functional materials in composition, physical and chemical properties. This is true for instance in the metallurgy and materials processing of metallic alloys and non-metalic material. However, it also applies in various fields of technical chemistry, such as the distillation of gasoline. Silicon technology and herewith the whole field of microelectronics depends on crystal growth based on phase transitions. Processing of thin film devices relies on phase transformations involving gaseous or plasma like phases into solids. The list of examples could be optionally extended even further.

6 Fundamental thermodynamic considerations

The variation of the state of existence of matter, i.e. phase transitions, and the conditions under which several modifications can even coexist in equilibrium are described by thermodynamics. Before treating this field in more detail later within in this section of the present tutorial it appears helpful and even necessary to define some basic terms and notions.

6.1 Definitions

- *Components*

  In a particular material systems in which more than one chemical element or species are involved it is important to define the notion of components. In the framework of chemical thermodynamics the term components refers to chemically independent elements or compounds within one system. As a consequence materials science and chemistry defines unary, binary, ternary, quaternary systems – and so on – depending on the number of components *C* involved. A unary system is for example pure iron Fe, one of the most important constituents of structural materials: steel. Below 912°C it crystallizes in the solid body centred cubic modification (α-Fe, ferrite), which is ferromagnetic up to

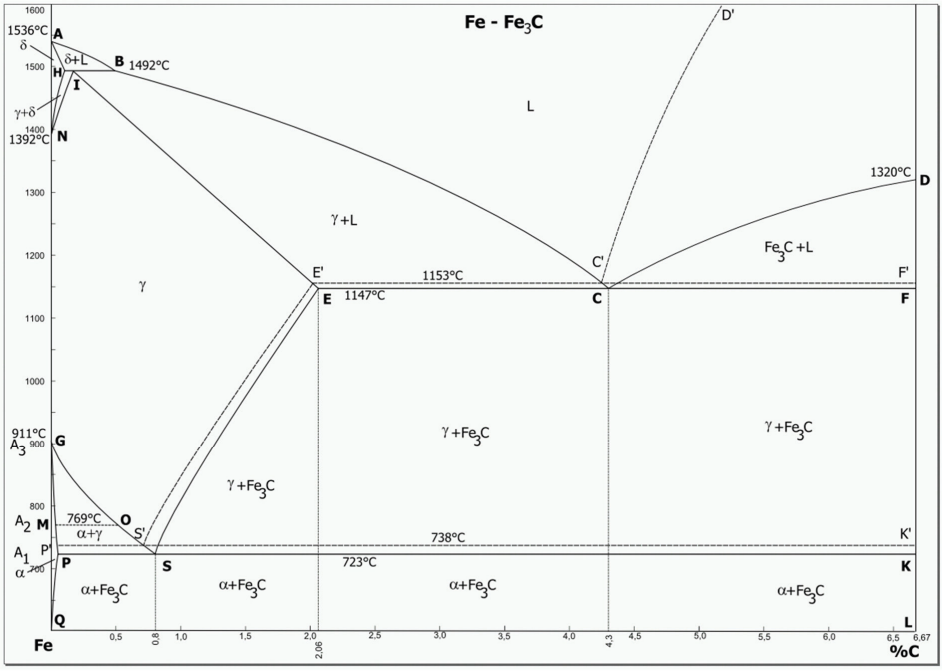


Fig. 13: *The isobar (p=const.) phase diagram Fe – C. The solid lines represent the stable version the broken lines the metastable one, that applies to cast iron alloys, directly solidified from the molten state, where graphite C may appear as a phase instead of cementite Fe_3C.*

770°C and paramagnetic beyond this temperature. Above 912°C the crystallographic structure changes to a face- centred cubic modification (γ-Fe, austenite) and at even higher temperatures exceeding 1394°C it converts again to a high temperature solid body-centred cubic modification that is stable up to 1538°C, the melting point of iron. At pressures above approximately 10 GPa and temperatures of a few hundred Kelvin or less, α-Fe changes into a hexagonal-closed packed structure, which is also known as the ε-phase. This variety of different solid modifications of one and the same element – in this case iron Fe – is called "*Allotropy*". Technically more important than pure iron are its alloys with carbon C and many other metallic elements (among which chromium Cr, manganese Mn, cobalt Co, molybdenum Mo, tungsten W, niobium Nb, vanadium V and several others). The simple Fe-C system on which most of them are based is a binary system, since Fe and C may form an intermediate compound Fe_3C, cementite – as already outlined in section 5.2. For this reason they are not chemically independent.

The complexity that the introduction of even relatively small amounts of a foreign element to a pure substance causes is expressed in the example of the so called Fe-C – diagram which depicts the stability regions of various constitutions of steels as a function of temperature (ordinate) and C-content (abscissa).

Considering back the former example of pure water H_2O of section 3 in comparison with the present case of the combination of iron Fe and carbon C reveals, that here the stoichiometry in contrary to Fe-C alloys is fixed. Water only exist in the inflexible molar ratio of two hydrogen atoms H and exactly one oxygen atom O. This relative molecular proportion does not change, even when water transforms from its solid structure into a liquid or if it vaporizes.

Both examples discussed in this subsection – involving two chemical elements each Fe and C on one side and H and O on the other – illustrate how important it is to specifically identify the number of components in thermodynamics.

- *Phases and phase boundaries: homogeneity vs. heterogeneity*

The question

"What defines a phase?"

isn't so easy to answer. A very short, simple and classical textbook version giving an explanation originates from John Willard Gibbs (1839 – 1903):

"A phase is a homogeneous part of an eventually heterogeneous system".

This characterization, however, strictly only applies if no external forces interact with the system under consideration. Intuitively, one would regard – referring again to the simple example of section 3 – the cloudless atmosphere during a sunny summer day as one single and homogeneous phase. A more rigorous consideration, however, reveals that this is not precisely correct, because of density fluctuations due to the gravitational field of the earth. A more suitable definition would therefore be:

"In a system consisting of several areas of space, in which the properties defining their state, respectively, only change steadily by very little variations and in which those areas are in contact with each other in a manner that the state characteristics change abruptly on a very small length scale at the borders separating them, the individual regions are designated as phases and their borders as phase boundaries."

This more accurate definition, however, also illustrates the problematic difficulty of exactly describing, what is a single homogeneous phase and what is a heterogeneous phase mixture, since it of course depends on what is considered as a small length scale. Milk, ink or blood macroscopically all appear as single phase fluids, but on the microscopic level, these colloids are in fact heterogeneous. Defining the chemical or physical uniformity of a certain phase is in fact – a little philosophical – *"a question of point of view"*: whether we investigate the materials properties and structure

- macroscopically by some physical measurements,
- optically with our eyes,
- through microscopic inspection with a simple light microscope,
- by observation with electron microscopy (SEM, TEM) even up to high resolution (HR-TEM),
- or: by the application of other methods. Like for instance Raman-spectroscopy that reveals respectively very tiny structural crystallographic variations even on the molecular level or on the level of the unit cell of a certain lattice symmetry.

6.2 The Gibbs phase rule

Around 1876 the American engineer John Williams Gibbs specified in a simple mathematically formulated rule, how many phases in an eventually multi-component system may coexist. It turns out that the degrees of freedom F, representing the choice of external physical or chemical parameter determining one specific state of mater, being pressure, temperature or composition or whatever else can be chosen freely, without changing the stability of coexisting phases. The deduction of this rule follows a strict thermodynamically treatment and will not be discussed in detail in the framework of the present tutorial – at the present stage at least. Some instructive simple examples have already been described in section 3 of the present chapter, where two cases were considered for unary systems: phase transitions in pure water H_2O and the liquefaction of a gas by compression.

For all fluid systems (liquids and solids), that in principle often show a relatively high compressibility (reduction of volume V by the application of external pressure p) the Gibbs phase law predicts that the maximum number of degrees of freedom F, for which an equilibrium of a certain number of phases P in a system consisting of a number of components C can be expected according to the laws of thermodynamics amounts to:

$$F = C - P + 2 \tag{7}$$

As will be outlined later in the framework of this text the number of maximum degrees of freedom F corresponds exactly to the number of variables of states that can be changed without altering the numbers of coexisting phases for a given number of independent components C. Coming back to the initial example of pure water (treated in section 3) – an unary system (C=1) – the maximum number of freedom for a single phase region (P=1) would be F=1-1+2, so exactly two. This can easily be appreciated from Fig. 2. In all two-dimensional phase regions for the single phases vapour, liquid or solid, of the p-T phase diagram both parameters pressure and temperature can be changed in certain limits without modifying the phase equilibrium. The situation changes if one refers to the freezing/melting line (equilibrium and phase boundary between ice and liquid water phase), to the condensation/vaporization line

(equilibrium and phase boundary between liquid water and gaseous phase) or to the resublimation/sublimation line (equilibrium and phase boundary between ice and gaseous phase). The number of coexisting phases increases to two, the resulting number of maximum degrees in freedom is reduced to one: phase coexistence can only be maintained if the change of one parameter – pressure or temperature – is counterbalanced by the adoption of the other one. Even more particular is the situation at the so called triple-point, where water ice and vapour coexist. The number of concurring phases for stability increases to three, leaving only a degree of freedom of zero: this state of matter is defined by one specific temperature and one precise value of pressure only.

In solid state systems the phase rule, deduced by Gibbs simplifies, since the low compressibility of condensed systems allows neglecting the influence of mechanical pressure. Consequently the formula now reads, as follows:

$$F = C - P + 1 \tag{8}$$

Without going to much into details, the example of the binary system iron – carbon (Fe – C) – as one case of a higher ordered and heterogenous systems should be reconsidered again. Pure Fe exists in its face-centred cubic modification (γ-Fe, austenite) between 912 °C and 1394 °C. Impressively, small amounts of carbon added, significantly change the phase boundaries in this system in that way that austenite is stable in a larger temperature range than without dissolved carbon C – at least up to a concentration of 0.8 at.-%. In this condensed binary system (C=2) two degrees of freedom can be chosen arbitrarily in certain limits without changing the phase stability of the single phase γ-Fe (austenite, P=1): temperature and composition (C-content). The influence of mechanical pressure was neglected here according to the Gibbs-phase rule of condensed systems.

It should be noted, however, that the simplification to exclude pressure is not admissible in geological sciences such as mineralogy, where the significantly enhanced values of mechanical pressure due to the presence of the earth crust have to be taken into account.

6.3 Order parameters

In almost all cases individual possible modifications or phases of matter differ in their symmetry or order. The physical parameter defining the differences of various phases of one system is therefore specified as *order parameter* ξ. It is used to describe the physical state distinguishing separate phases during a phase transition and represents a measure for the degree of order. Depending on the exact nature of the transition the ordering parameter can change discontinuously or steadily. Generally, the order parameter takes a value of 0 for the phase of higher symmetry and a value non equal to 0 for the one with lower symmetry. An intuitive example was already presented in the intermediate part of this tutorial (Fig. 4) on the different state forms of matter, where the order decreased from the solid, over the liquid and next the gaseous to finally the plasma state. A possible ordering parameter describing this sequence of phase transitions could be for example the density ρ that decreases with increasing degree of disorder. For such transitions that involve phases with a different state forms of matter (gas, liquid and solid), the ordering parameter – here density ρ – changes abruptly. On the contrary magnetic polarization M or electric polarizations P (refer to section 5.2) of a solid occur continuously, as a result of the change of external conditions, like temperature T, pressure p, magnetic field strength H or electric field strength E.

Other illustrating cases of order parameter are added in the list shown in table below.

Phases involved		**Transition Process**	**Order parameter**
Gas	Liquid	Vaporization, Condensation	Density ρ
Liquid	Solid	Melting, Solidification, Crystallization	
Gas	Solid	Resublimation, Sublimation	
Liquid	Liquid	Unmixing	Composition (Concentration c)
Solid	Solid	Decomposition	Composition (Concentration c)
		Ordering	Long range order parameter
		Change in crystal structure	Distortion, lattice symmetry
Paramagnetic	Ferromagnetic or ferrimagnetic	Magnetization	Spontaneous magnetization M
Paraelectric	Ferroelectric		Electrical polarization P

In many cases of solid to solid phase transitions different phases often differ in not only one single order parameter but in several ones. Examples are ferromagnetic or ferroelectric transformations, where it is not solely magnetization or electrical polarization that change but also mechanical distortion due to electro- or magnetostriction.

6.4 Thermodynamical functions, process parameters and equilibrium

- *Thermodynamical functions and potentials*

  A very fundamental and essential law in natural sciences teaches that a system, of whatever of kind it is alike, is stable and in equilibrium, when its energy is minimal, when it is as small as possible. The term *equilibrium* means that the system is then in a state, that will not change anymore and remain as it is, unless the external physical or chemical conditions are modified. This concept originally derived from mechanics arises from the fact, that if the driving force for a certain process such as motion, a chemical reaction or a phase transition vanishes, the system under consideration will be invariant. Since force

is formally and mathematically in generally a first derivate of energy this translates into the fact that energy should be minimized to realize a stable situation. From this context it becomes evident, that a judgement on the stability regions of matter in dependence of external physico-chemical parameters, like pressure, temperature, applied magnetic, electric even or even elastic fields cannot be described completely quantitatively in terms of the single order parameters ξ alone, which in fact are only specific and particular experimentally accessible *"markers"* for observable modifications of the materials behaviour or state. A full treatment requires the description of the inner energy U stored in a certain system in the framework of thermodynamics, correlating process quantities such as heat Q and mechanical work W to so called state functions, that energetically quantify a certain state.

Independently of in which kind of state a particular material system is, the inner energy U represents the summation of all kinetic and potential energy contributions, resulting from the external transfer of heat Q and or mechanical work W stored in that system. They include vibrational and eventually rotational parts of elementary particles like atoms, ions or molecules and their eventual arrangements arising from heating above a temperature above the absolute zero point of temperature: 0 K (Kelvin). This is exactly stated by the zeroth law of thermodynamics:

"If two systems are in thermal equilibrium independently with a third system, they must be in thermal equilibrium with each other."

In any case, the first law of thermodynamics, teaches that energy can never be lost. This fact implies that, whatever kind of energy portion is transferred to a certain system – be it a partial differential of mechanical work δW or caloric heat δQ or of any other source is stored as a complete differential:

$$\mathrm{d}U = \delta Q + \delta W \tag{9}$$

The second law of thermodynamics postulates that thermal energy cannot be transformed at any rate in any another form of energy. This means that a certain portion of internally stored energy cannot be explored by any means. Or expressed in the other original words: there is no variation of matter, which results in the simple transaction of heat Q from of one body of lower temperature to one of higher temperature. This statement at the first sight seems obscure, but it can be illustrated by a simple "Gedanken" experiment: if a certain substance spontaneously dissolves in another substance (taken a droplet of ink immersing in some specific amount of water), because it's a energetically more favourable situation, generally there is no force to push the system back to un-mixing. Once a stable mixture is formed it will remain as it is.

This kind of irreversible contribution did bring up the term of *"entropy" S*. Intuitively, the entropy S can be regarded as a measure for the degree of disorder in a certain system: the higher its value, the larger is the disorder. This concept describes that spontaneously occurring processes are irreversible.

Hermann von Helmholtz (1821 – 1894) introduced a novel concept, only accounting for the free parts of energy that can be transferred and defined, what later was named to his honour the free Helmholtz-energy F:

$$F = U - T \cdot S \tag{10}$$

An alternative formulation of the energy content of a certain system is the quantity of the enthalpy H:

$$H = U + p \cdot V \tag{11}$$

However the most pragmatically used variable nowadays is the free Gibbs-enthalpy G:

$$G = H - T \cdot S \tag{12}$$

- *State variables, intensive and extensive properties*

  The further treatment needs a further more systematical distinction between intensive and extensive parameters describing a certain thermodynamical material system. *Intensive properties* such as pressure p, temperature T, electrical field strength E or magnetical field strength H do not correlate with the size of the system under consideration. On the other hand, *extensive properties*, such as volume V, inner energy U, free Helmholtz-energy F, enthalpy H, free Gibbs-enthalpy G, entropy S, electrical polarization P and magnetization M depend on the size of the represented system.

  The product of intensive and extensive properties results in an energy.

- *Classification of phase transitions*

  According to the Austrian physicist Paul Ehrenfest (1880 – 1930) phase transitions can be categorized according to their *"order" n*. In this context systems are described by thermodynamical potentials, like the free Helmholtz-energy F or free Gibbs-enthalpy G in dependence of state variables, such as for instance temperature T, pressure p. The order n of a phase transition then defines where the thermodynamical potential becomes unsteady and at which degree n of its derivative a singularity is observed.

  First order transitions represent discontinuous phase modifications, because the first derivative of the thermodynamical potential, for example

$$V = -\frac{\partial G}{\partial p} \text{ or } S = -\frac{\partial G}{\partial T} \tag{13}$$

  and the order parameter ξ (such as for instance density ρ) reveal a singularity. Therefore, latent heat is consumed or released during first order transitions, since local infinitesimal fluctuations in the state of the system cannot expand spontaneously but require a certain activation energy for the transformation. Typical examples for first order phase transitions, treated so far in the present tutorial, are the modifications of the state form of matter and crystallographic transitions. They consume energy for the formation of nuclei that are able to grow spontaneously, since new interfaces between the coexisting phases have to be established (section 7). This nucleation energy is compensated by lowering the free Gibbs-enthalpy of the system, for instance. Correspondingly, hysteresis effects such as undercooling and overheating are characteristic for first order transitions: more generally this means that the point where they occur can differ depending on whether a state variable is in- or decreased.

  Second order transitions, on the other hand, are continuous processes. The first derivative of the thermodynamical potential (e.g. equation 13) is unsteady at the transformation point but does not show any singularity as in the case of transitions of the order one. Also the order parameter, for instance the magnetization M or electrical polarization P change continuously in dependence of intensive properties such as temperature T, pres-

sure p, magnetic field strength H or electric field strength E at the transition point. No latent heat and generally no interfacial extra energy are involved during a transition of second order. Phase transitions based on the mechanism of spontaneous un-mixing (spinoidal decomposition) may be added to this category.

- *Landau-Theory*

  According to the so-called Landau-Theory [9], developed by the Soviet theoretical physicist Lev Davidovich Landau (1908 – 1968, Nobel prize in physics 1962), second order phase transitions are described in the framework of a phenomenological model, relying on group theory. In this formalism the thermodynamical potential is expressed in function of the order parameter ξ (e.g. magnetization M or electric polarization P), designating properties that are accessible and determinable experimentally on a macroscopic scale, in the vicinity of the phase transition point. Generally, the thermodynamical potential, for instance free Gibbs-enthalpy is developed and expressed in the form of a polynominal series expansion. The model usually applies in an approximate way to phase transitions, where certain symmetrical elements are broken. Since the Landau description of phase transitions represents a phenomenological theory, it only accounts for macroscopically observable material properties – as for instance electric polarization P of a ferroelectric material. The theory therefore cannot reveal any information on microscopic source for a certain material behavior.
  The coefficients of the polynominal expansion are derived macroscopically by experiments. For this reason the Landau-model only represents a mean field theory, since it averages on microscopic interactions. Fluctuations of the order parameter around its equilibrium state are not considered.

  Generally the thermodynamic potential (e.g. the free Gibbs-enthalpy G) in the Landau-formalism takes the form – written in an exemplarily chosen way:

$$G(T, p, \xi) = G_0(T, p) + \frac{A(T, p)}{2} \cdot \xi^2 + \frac{B(T, p)}{4} \cdot \xi^4 + \ldots \tag{14}$$

  Principally, the coefficients G_0 A, B and so on in this development depended on external thermodynamic variables.

7 Kinetic aspects: nucleation and growth

Thermodynamics only describes and predicts, as shown in the previous part (section 6) of this tutorial, what kind and in which number phases in a particular material system are stable and eventually in equilibrium under certain external conditions, such as pressure, temperature and chemical composition, for instance. However, no predictions on how distinct phases transform into each other can be made. Generally two different possible scenarios can be distinguished in this context regarding the stability of the initial phase through the variation of external physical conditions:

- The initial phase becomes completely *unstable*. In this first case the system will transform or decompose spontaneously in a way that a new phase continuously grows by infinitesimal small fluctuations out of the initial phase (e.g. negative or uphill-diffusion, spinoidal decomposition – refer to section 5.2, page A5 – 8). Even if submicronic actual-

ly thermodynamically unstable nuclei are formed by coincidence, because the interaction forces driving the creation of the new phase predominate the arbitrary thermal motion of the elemental particles transformation will occur. In the case of spinoidal decomposition, for instance, this means that against the rules of diffusion (Fick's first and second law) individual particles will not move in a way that concentration gradients are equalized but in an inverted manner. Concentration fluctuations gradually reinforce spontaneously. The reason for this behavior is that the attractive forces of two different components constituting an initial phase before transformation are stronger under the changed external physico-chemical conditions than before.

- The initial phase remains simply *metastable* and continues to exist even under the changed external circumstances (by undercooling or overheating beyond its stality point). In this second case this metastable phase transforms into a new more stable phase through nucleation and growth. Since the initial phase still reveals a certain stability infinitesimal small fluctuations are not sufficient to form the new phase. Only if larger fluctuations occur, sufficient to develop overcritical nuclei that are capable of growth, the transformation starts to take place. In a later stadium, when the initial phase already nearly disappeared completely, the transformed areas of the new phase can further grow on the expense of other smaller ones that continuously decrease in number and finally vanish completely (*Ostwald ripening*).

The case of nuclei formation and growth will be considered in more depth – whereas the case spinoidial decomposition, is left for the interested reader to follow more intensively in the online available literature or in classical libraries.

Assuming a simple phase α, that would transform into a second phase β because some external chemical or physical change in circumstances, a certain energetic gain could be expected for instance if this phase is undercooled in a metastable state, in which thermodynamics would formally predict its inexistence for example by simply lowering temperature. Because of the reduction of volumetric free Gibbs-Enthalpy ΔG_V (free Gibbs-enthalpy G per volume V) in this thermal range of undercooling additional energy is available to establish the interfacial energetic part – ignored so far in purely thermodynamical considerations – between the mother phase and the emerging new nucleus.

For the geometrically simplest case of a assumed spherical nucleus with a radius r the gain of volumetric amounts to:

$$\frac{4\pi}{3} \cdot r^3 \cdot \Delta G_v \tag{15}$$

On the other hand the creation of a new interface of the spherical nucleus to be formed requires and consumes energy, since the bonding interactions at the surface are different to the ones in the volume:

$$4\pi \cdot r^2 \cdot \sigma \tag{16}$$

σ represents the specific surface energy between the two coexisting phases. The energetical balance of the system can then be expressed as:

$$\Delta G = \frac{4\pi}{3} \cdot r^3 \cdot \Delta G_v - 4\pi \cdot r^2 \cdot \sigma \tag{17}$$

In order to form a critical nucleus of critical radius r_c able to grow spontaneously the first derivative of this term over the radius r has to be zero:

$$\frac{\partial \Delta G}{\partial r} = 4\pi \cdot r_c^2 \cdot \Delta G_v - 8\pi \cdot r_c \cdot \sigma = 0 \tag{18}$$

This results in an expression for the critical nucleus radius r_c

$$r_c = \frac{2 \cdot \sigma}{\Delta G_v} \tag{19}$$

and for the activation energy ΔG_c necessary for its formation:

$$\Delta G_c = \frac{16\pi \cdot \sigma^3}{4 \cdot \Delta G_V} \tag{20}$$

This treatment only considers the homogeneous nucleation. Generally however, phase transformation start to occur heterogeneously at existing interfaces of defects within a material, because this is energetically more favourable.

The time dependence (kinetics) of phase transitions can approximately be described by the so-called *Johnson-Mehl-Avrami-Kolmogorov* – equation (JMAK – equation) at isothermal conditions (T = const.). This theory [10] describes the process for microstructural transformations using two properties, treating both these properties like chemical reaction rates: a constant nucleation rate and constant growth rate of already existing nuclei. The JMAK-equation is an important and fundamental basis for the calculation of so called TTT-diagrams (Time – Temperature – Transformation – diagrams. It predicts the volume fraction of a newly formed phase, depending on time assuming that during a phase transition constantly but not simultaneously new spherical nuclei are created randomly in space and grow. At the same time as these nuclei grow, new nuclei are formed.

The proportion f of the transformed phase out the entire initial volume of the system is then expressed as:

$$f = 1 - \exp(-\frac{\pi}{3} \cdot N \cdot v^3 \cdot t^4) \tag{21}$$

This relationship holds for short as well as long transformation times t and also for small as large fractions f of the newly created phase out of its matrix. N describes the number of nuclei formed per time unit (nucleation rate determined by the activation energy ΔG_c for the formation of critical nuclei) and v (growth rate). For short times t the expression simplifies to

$$f = \frac{\pi}{3} \cdot N \cdot v^3 \cdot t^4 \tag{22}$$

due to the approximation $1 - \exp(x) \approx x$ for small values of x far below 1.

The equation for short times can in a simplified manner explained as follows: the number of nuclei with time increases according $N \cdot t$ and the individual volume of a nucleus growth with $4\pi/3 \cdot (v \cdot t)^3$, since the radius r increases with time t according to $v \cdot t$.

References

[1] V. Radmilovic, C. Ophus, E.A. Marquis, M.D. Rossell, A. Tolley, A. Gautam, M. Asta, U. Dahmen : *Nature Materials* **10** (2011) 710-715.

[2] S.D. Bakshi, A.Leiro, B. Prakash, H.K.D.H Bhadeshia: *Material Science and Technology* **31** (2015) 1735-1744.

[3] Y.F. Shen, X.X. Li, X. Sun, Y.D. Wang, L. Zuo: *Materials Science and Engineering A* **552** (2012) 514-522.

[4] A.K. Chakrabarty, K.T. Jacobs: *Journal of Phase Equilibria and Diffusion* **34** (2013) 267-276.

[5] C. Pithan: "Permanentmagnetische Sinterwerkstoffe auf Fe – Nd – B – Ga-Basis mit Nd-armer Korngrenzenphase", *Ph.D.-Dissertation*, University of Stuttgart (1993).

[6] C.-L. Jia, S.-B. Mi, K. Urban, I. Vrejoiu, M. Alexe, D. Hesse, *Nature Materials* **7** (2008) 57-61.

[7] http://www.doitpoms.ac.uk/miclib/micrograph.php?id=199.

[8] A. Hubert, R. Schäfer: "Magnetic domains – the analysis of magnetic microstructures", Springer, Berlin, Heidelberg, New York (2009) 510.

[9] L. Landau, *Ukranian Journal of Physics* **53** (2008) 25-35.
(English translation of the original version published at: *Zh. Eksp. Teor. Fiz.***7** (1937) 19–32.

[10] M. C. Weinberg, D. P. Birnie III, V. A. Shneidman, *Journal of Non-Crystalline Solids* **219** (1997) 89–99.

A 6 Physics and Chemistry of Redox Processes

Rotraut Merkle
Max Planck Institute for Solid State Research
Stuttgart, Germany

Contents

1 Introduction - redox reactions **2**

2 Oxidation reactions and oxygen exchange **3**
2.1 Thermodynamics - Ellingham diagrams 3
2.2 Stoichiometry relaxation and related experiments 4
2.3 Kinetics and mechanism of the oxygen exchange surface reaction 7
2.4 Kinetics of oxide scale formation 12

3 Redox processes in electrochemical cells **14**
3.1 Electrochemical cells at open circuit 14
3.2 Cells generating current or acting as electrochemical pump 17
3.3 Electrode kinetics 18
3.4 Stoichiometry polarization, Wagner-Hebb experiments 19

4 Processes driven by voltage load or *T* gradients **21**
4.1 Kinetic demixing 21
4.2 Thermodiffusion 22

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction – redox reactions

Reactions in which electrons are transferred between the reaction partners, and in which correspondingly the oxidation states of atoms change, are denoted as redox processes. Some examples:

- The oxidation of hydrogen occurs as a homogeneous reaction in the gas phase

$$H_2 + 1/2\ O_2 \rightleftharpoons H_2O \tag{1}$$

  and at elevated temperature also the reaction product H_2O is gaseous.

- The oxidation of metals such as Ti converts one solid phase (Ti) into a different solid phase (TiO_2)

$$Ti + O_2 \rightleftharpoons TiO_2 \tag{2}$$

- At temperatures above zero K, any oxide exhibits a small but nonzero deviation from the perfect stoichiometry (see chapter A3). When temperature or oxygen partial pressure pO_2 are changed, this nonstoichiometry δ has to adjust, e.g.

$$TiO_{2-\delta} + \varepsilon/2\ O_2 \rightleftharpoons TiO_{2-\delta+\varepsilon} \tag{3}$$

  and when the perturbation is not too large, the material remains single-phase.

Redox reactions can proceed by direct chemical reaction between the reaction partners. They can often also be carried out in electrochemical cells, in which an electrolyte (pure ion conductor) is used to split the reaction into electron- and ion-transfer steps (see section 3), and which allows one to directly convert chemical energy into electrical energy and vice versa (fuel cells, batteries, electrolysers). Electrochemical cells can also transform "chemical information" (e.g. the presence of a gas such as CO_2) into an electrical signal, acting as a sensor. In the opposite direction, redox reactions in electrochemical cells can be used to create "chemical information" (e.g. deposition of metal clusters, filaments) by applying a voltage or current.

For electrochemical devices, redox processes are important in various functions. They may be involved in the materials preparation steps (typically at very high temperatures for oxide and ceramic devices). When oxides are not deliberately sealed from the atmosphere, their nonstoichiometry slowly equilibrates with the pO_2 at the lower operation temperature which modifies the point defect concentrations and thus the materials properties, see chapter A3. When the operation temperature is so low that the oxygen exchange reaction is kinetically frozen in, the adjustment of the nonstoichiometry at high T combined with quenching can be used to tune the materials properties (e.g. T_c in $YBa_2Cu_3O_{6+\delta}$ superconductors).

Redox processes are an integral part of the functioning of electrochemical devices (e.g. the electrode kinetics of fuel cells, batteries, and gas sensors) but are also important for the long-time stability of devices (e.g. a low oxidation rate is required for metallic interconnects in solid oxide fuel cells SOFC). Thus, a detailed understanding of the thermodynamics and kinetics of such redox processes is required.

The present chapter discusses redox processes initiated by several driving forces. We start with the chemical driving force (typically a difference in the chemical potential of oxygen). In section 2.2, the overall stoichiometry relaxation kinetics and closely related experiments will be discussed, while 2.3 focuses on the surface reaction mechanism. Section 2.4 deals with

oxide phase formation in response to the chemical driving force. Section 3 describes electrochemical cells without and with current load. Further ion transport processes are discussed in section 4: demixing of solid solutions in electrical and pO_2 gradients, and ion transport driven by temperature gradients. More detailed discussions of the topics treated in this chapter can be found in textbooks and overview articles such as [1,2,3,4,5,6,7,8,9].

2 Oxidation reactions and oxygen exchange

2.1 Thermodynamics - Ellingham diagrams

The thermodynamics of chemical reactions such as metal oxidation

$$a\mathrm{M} + 1/2\ \mathrm{O_2} \rightleftharpoons \mathrm{M}_a\mathrm{O} \tag{4}$$

is described by the change of the standard Gibbs free enthalpy

$$\Delta_r G^0 = \Delta_r H^0 - T\Delta_r S^0 \tag{5}$$

$$= H^0(\mathrm{M}_a\mathrm{O}) - aH^0(\mathrm{M}) - 0.5H^0(\mathrm{O_2}) - T\left[S^0(\mathrm{M}_a\mathrm{O}) - aS^0(\mathrm{M}) - 0.5S^0(\mathrm{O_2})\right] \tag{6}$$

A useful representation is the "Ellingham diagram" which shows $\Delta_r G^0$ as function of T for several reactions (Fig. 1a). Since to good approximation the standard enthalpy of reaction $\Delta_r H^0$ and entropy of reaction $\Delta_r S^0$ are T-independent (unless phase transformations occur), $\Delta_r G^0(T)$ yields essentially straight lines with axis intercept $\Delta_r H^0$ and slope $\Delta_r S^0$. The standard entropy of oxidation reactions is usually negative because gaseous O_2 molecules are converted into a solid oxide, losing their translational and rotational degrees of freedom. This leads to a positive slope in the Ellingham diagram. Less noble metals (which have a higher driving force for oxidation) appear in the lower part of the Ellingham plot (cf. MgO vs. NiO, ZrO_2 vs. TiO_2).

For each T, there is exactly one pO_2 under which metal and oxide phases coexist (Gibbs phase rule). This pO_2 is obtained from the mass action constant of the oxidation reaction (R = universal gas constant; the activities of the pure solid phases are constant and therefore included in K_{ox}):

$$K_{ox} = (pO_2)^{-1/2} = e^{-\Delta_{ox}G^0/RT} \tag{7}$$

These values are indicated in Fig. 1b. Metals which form several oxides (e.g. Cu_2O and CuO, dashed lines in Fig. 1) exhibit one coexistence line metal/low-oxidation state oxide determined by eq. (7), and one between the two different oxides, determined by the mass action constant for the interconversion reaction such as $1/2\ Cu_2O + 1/4\ O_2 \rightleftharpoons CuO$. The stable regions for Cu, Cu_2O and CuO are indicated by the labels in boxes in Fig. 1b

The gas-phase reactions $H_2 + 1/2\ O_2 \rightleftharpoons H_2O$ and $CO + 1/2\ O_2 \rightleftharpoons CO_2$ are often used to establish well-defined low oxygen activities. The resulting pO_2 values depend on the H_2, H_2O and CO, CO_2 ratio as well as on temperature. For example, a 1:1 mixture of H_2 and H_2O yields $\log(pO_2/\mathrm{bar}) = -15$ at 900 K and -8 at 1500 K. As long as this pO_2 is within the single-phase region of an oxide, the applied pO_2 can be used to tune the point defect concentrations of the oxide (see chapter A3).

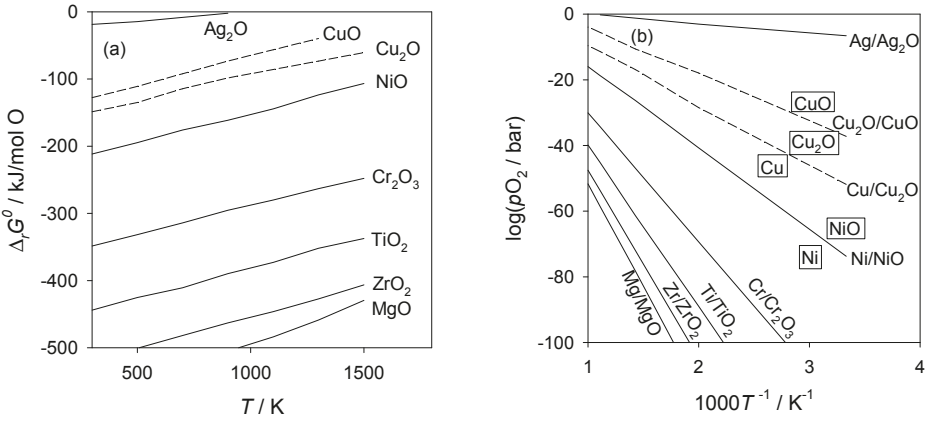


Fig. 1: *(a) Gibbs free enthalpy of the reaction aM + 1/2 O_2 ⇌ M_aO, normalized per mol O in the oxide. (b) Metal/oxide coexistence lines; for the Cu-O system also the Cu_2O/CuO coexistence line is given. Data taken from* [10].

2.2 Stoichiometry relaxation and related experiments

As long as an oxide is kept within the single-phase range, a change of T and pO_2 leads to an adjustment of its nonstoichiometry according to

$$M_aO_{1-\delta} + \varepsilon/2\ O_2 \rightleftharpoons M_aO_{1-\delta+\varepsilon} \tag{8}$$

This general reaction holds for oxygen deficiency (positive values of δ) as well as oxygen excess (negative δ), and for all possibilities of realizing this nonstoichiometry (oxygen vacancies or interstitials, cation vacancies or interstitials). It is a redox process because oxygen is converted from O_2 (oxidation state 0) to oxide ions in the lattice (formal oxidation state -2; the real ion charge is typically a bit lower owing to partially covalent bonding, see e.g. [11]). In parallel, conduction electrons are consumed or electron holes created. Even if the electronic conductivity of the oxide is low, some transport of electronic carriers has to take place in parallel to the incorporation of oxide ions to fulfill the electroneutrality condition (in this sense even MgO is a "mixed conductor"). The kinetics of this stoichiometry relaxation is determined by one of several processes (see [1,4,6] for more details) as indicated in Fig. 2:

- surface exchange reaction $1/2\ O_2 + V_O^{\bullet\bullet} \rightleftharpoons O_O^x + 2h^\bullet$ (Kröger-Vink notation, see chapter A3), which transforms O_2 molecules into oxide ions in the first layer of the oxide. When the surface reaction limits the overall relaxation kinetics, the flux of O is given within the framework of linear irreversible thermodynamics by

$$j_O = -k_O^\delta \delta c_O \tag{9}$$

where k^δ is the effective rate constant which is discussed in more detail in section 2.3, and δc_O is the step-like concentration change of O at the surface (corresponding to the drop in μ_O between the new equilibrium value and the actual value within the oxide drawn in Fig. 2). This equation has the same form flux = transport coefficient × concentration gradient as Fick's first law (eq. (10) below), with the transport coefficient being k^δ and the concentration gradient being condensed into the step δc_O at the surface.

- bulk chemical diffusion of oxygen which occurs by coupled migration of ionic (e.g. oxygen vacancies $V_O^{\bullet\bullet}$) as well as electronic defects (e.g. electron holes $h^{\bullet}$), cf. chapter A4. When this limits the overall kinetics, the O flux is determined by the chemical diffusion coefficient D^{δ}

$$j_O = -D_O^{\delta} \nabla c_O \tag{10}$$

- O transfer across blocking grain boundaries (cf. chapter A4; boundary blocking for ionic and/or ionic defects which are both required for chemical diffusion). The transfer across an individual blocking grain boundary is described by

$$j_O = -k_{O,gb}^{\delta} \delta c_{O,gb} \tag{11}$$

where $k_{O,gb}^{\delta}$ is the effective rate constant for the grain boundary transfer and $\delta c_{O,gb}$ the respective concentration step.

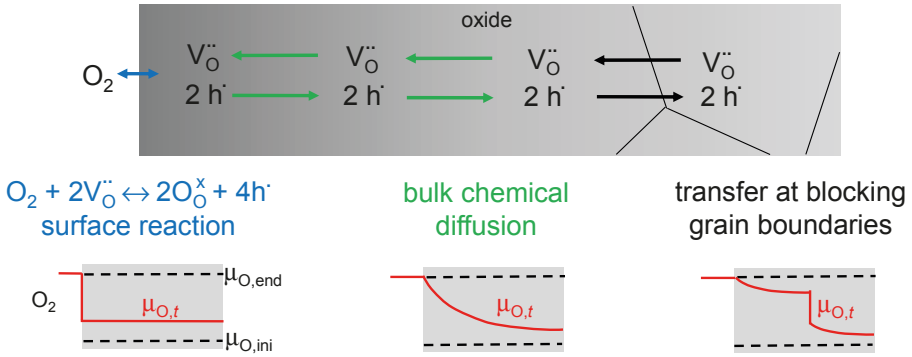


Fig. 2: *Stoichiometry relaxation of an oxide after a stepwise increase of pO_2. The overall kinetics can be determined by surface reaction, bulk chemical diffusion of oxygen, or transport across blocking grain boundaries. The drawings at the bottom show the profile of the oxygen chemical potential for each of these processes being limiting.*

The surface reaction is limiting when the relation $k \cdot l < D$ holds [12], where l is half the sample thickness. This relation is applicable for chemical, tracer and electrochemical oxygen exchange by inserting the respective k and D. In the surface reaction limited regime, the concentration profiles are flat within the sample and the change in μ_O is concentrated to the step at the surface as indicated in the insets in Fig. 2. Vice versa, the oxygen exchange is diffusion limited for $k \cdot l > D$, i.e. thick samples and/or fast surface reaction. Then during the relaxation c_O and μ_O exhibit sloping profiles over the whole sample thickness.

For 1-dimensional concentrations profiles (sample = thin plate) and small driving forces (pO_2 changed by << 1 order of magnitude) the time evolution of the defect concentrations $\bar{c}$ averaged over the whole sample is given by the following equations. For surface limitation,

$$\frac{\bar{c}_t - \bar{c}_\infty}{\bar{c}_{t=0} - \bar{c}_\infty} = e^{-k^{\delta} t / l} \tag{12}$$

where l is half the sample thickness. For diffusion control, the short- and long-time expressions are

$$\frac{\bar{c}_t - \bar{c}_{t=0}}{\bar{c}_\infty - \bar{c}_{t=0}} \approx \frac{4}{\pi^{3/2}} \sqrt{\frac{t}{\tau_D}} \tag{13}$$

$$\frac{\bar{c}_t - \bar{c}_\infty}{\bar{c}_{t=0} - \bar{c}_\infty} = \frac{8}{\pi^2} e^{-t/\tau^\delta} \tag{14}$$

where $\tau_D = L^2 / (\pi^2 D^\delta)$ with L = full sample thickness. In the regime of mixed surface and diffusion control, k and D can be extracted by fitting the averaged defect concentration to [13]

$$\frac{\bar{c}_t - \bar{c}_{t=0}}{\bar{c}_\infty - \bar{c}_{t=0}} = 1 - \sum_{j=1}^{\infty} \frac{2L_\beta^2 \exp(-\beta_j^2 D^\delta t / l^2)}{\beta_j^2 (\beta_j^2 + L_\beta^2 + L_\beta)} \tag{15}$$

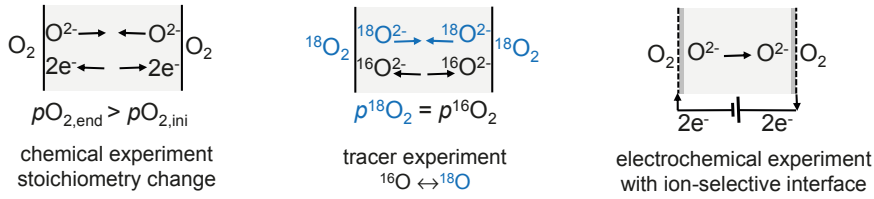
with $L_\beta = l \cdot k^\delta / D^\delta$, $\beta_j \tan \beta_j = L_\beta$ and l = half sample thickness. This can lead to larger uncertainties for k and D because the differences in concentration profiles are quite subtle. Space-resolved measurements such as ^{18}O profiles in the sample recorded by secondary ion mass spectrometry (SIMS [14], see chapter A4) or from space-resolved optical spectroscopy [15] allow for a better distinction between k and D contributions.

Oxygen exchange between gas phase and solid does not only occur in the "chemical experiment", i.e. change of pO_2 with corresponding stoichiometry change, but also in the isotope exchange "tracer" experiment and in the electrochemical experiment. These three closely related experiments are summarized in Fig. 3. While a change of oxygen potential μ_O and the corresponding ∇c_O is the chemical driving force for the first type of experiment, for the tracer experiment pO_2 is kept constant and only $^{16}O_2$ replaced by $^{18}O_2$. Thus, the driving force is purely entropic (a gradient in the isotope fraction ∇c_{18O}), and the oxide does not undergo a net stoichiometry change. In the electrochemical experiment, with the help of an ion-selective interface (a layer of electrolyte material) a purely ionic current is drawn through the mixed conducting oxide, with the electrical potential gradient $\nabla \phi$ as driving force. Also in this case the stoichiometry of the oxide remains unchanged, and correspondingly $k^* = k^q$ and $D^* = D^q$. The measurement of the surface kinetics of SOFC cathodes by impedance spectroscopy falls into this group, the obtained surface reaction resistance is $\propto 1/k^q$ (see e.g. [16]). The flux equations for diffusion and surface controlled kinetics are summarized in Fig. 3.

For the diffusion-limited case, the relations between D^δ, D^* and the defect diffusivity $D_{V_O^{\cdot\cdot}}$ were already discussed in chapter A4. They are included in Fig. 3 to emphasize the similarity of the expressions for k and D. For both surface as well as diffusion limitation, the appearance of the "thermodynamic factor" w_O is directly related to the fact that the stoichiometry is changed in the chemical experiment. w_O can be calculated when the oxygen stoichiometry is known as function of pO_2:

$$w_O = \frac{c_O}{RT} \frac{\partial \mu_O}{\partial c_O} = \frac{\partial \ln pO_2}{2 \partial \ln c_O} \propto \frac{1}{C^\delta} \tag{22}$$

For oxides with a small concentration of electronic carriers or redox-active centers such as $SrTiO_3$, w_O may take large values in the range of 10^5 [17], while for SOFC cathode materials it is much smaller, often $\approx 10^2$ [18]. This directly corresponds to the much larger "chemical capacitance" C^δ [19] of the latter materials.



diffusion limited:

$$j_O = -D^{\delta} \cdot \nabla c_O \qquad D^{\delta} \approx \frac{w_O}{c_O}[V_O^{\bullet\bullet}] \cdot D_{V_O^{\bullet\bullet}} \quad (16)$$

$$j_O = -D^{*} \cdot \nabla c_{18O} \qquad D^{*} \approx \frac{1}{c_O}[V_O^{\bullet\bullet}] \cdot D_{V_O^{\bullet\bullet}} \quad (17)$$

$$j_O = -D^{q} c_O \frac{2F\nabla\phi}{RT} \qquad D^{q} \approx D^{*} \quad (18)$$

surface limited:

$$j_O = -k^{\delta} \cdot \delta c_O \qquad k^{\delta} \approx \frac{w_O}{c_O} \Delta x \cdot \Re_0 \quad (19)$$

$$j_O = -k^{*} \cdot \delta c_{18O} \qquad k^{*} \approx \frac{1}{c_O} \Delta x \cdot \Re_0 \quad (20)$$

$$j_O = -k^{q} c_O \frac{2F\nabla\phi}{RT} \qquad k^{q} \approx k^{*} \quad (21)$$

Fig. 3: *Three exchange experiments: chemical, tracer, electrochemical. The equations for k and D are derived assuming $\sigma_{eon} > \sigma_{ion}$ (e.g. for a SOFC cathode material), and that the same surface reaction mechanism applies for all three experiments. Square brackets denote defect concentrations, the thermodynamic factor w_O is given by eq. (22) below. $\Re_0$ is the equilibrium exchange rate of the surface reaction, Δx is the spatial distance of 1-2 Å involved in the surface reaction between gas phase and first layer of the oxide.*

The expressions for D^{δ}, D^{*} as well as k^{δ}, k^{*} comprise a contribution which corresponds to a reciprocal capacitance (w_O/c_O for the chemical experiment, $1/c_O$ for tracer, electrochemical). The other part represents a reciprocal resistance ($[V_O^{\bullet\bullet}] \cdot D_{V_O^{\bullet\bullet}} \propto \sigma_{ion}$ for diffusion, $\Delta x \cdot \Re_0 \propto$ equilibrium exchange rate for the surface reaction). When the resistive part is identical, a larger capacitive part leads to a lower diffusion coefficient or effective surface rate constant because more atoms have to be exchanged upon a given external stimulus. The detailed derivation of eqs. (19-21) can be found in [1,4,6].

2.3 Kinetics and mechanism of the oxygen exchange surface reaction

The oxygen exchange reaction which converts gaseous O_2 into lattice oxygen in the first layer of the oxide is a multistep reaction that proceeds via chemisorption (which typically is fast), dissociation and incorporation steps as schematically shown in Fig. 4a. On ionic solids, oxygen typically forms charged adsorbates such as superoxide (O_2^-) or peroxide molecules (O_2^{2-}). Subsequently, the O-O bond has to be dissociated, which can occur without or with assistance by $V_O^{\bullet\bullet}$. Finally, the adsorbed atomic O^- has to be incorporated into a $V_O^{\bullet\bullet}$. The encounter of O^- and $V_O^{\bullet\bullet}$ can be achieved by migration of O^- and/or $V_O^{\bullet\bullet}$. The task is to identify the fastest of several parallel pathways, and within this the slowest step - the rate determining step (rds) which determines the overall reaction rate. For this, we first have to derive the equilibrium exchange rate $\Re_0$ for a hypothesized mechanism. Then, its dependence on pO_2 and defect

concentrations can be compared to experimental results. However, several hypothesized rds can lead to the same expression for $\Re_0$, which means that such cases cannot be distinguished by kinetics measurements alone.

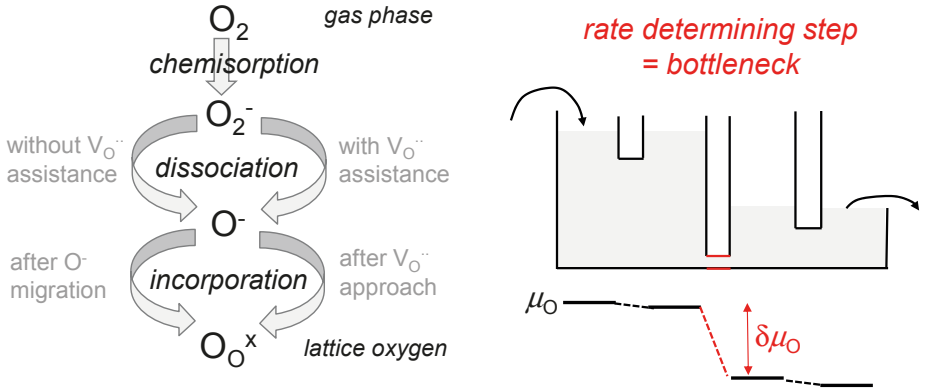


Fig. 4: *(a) Schematic: oxygen exchange is a multistep reaction, and can comprise parallel branches, e.g. dissociation without or with assistance of an oxygen vacancy. (b) Concept of the rate determining step: one step has a much lower exchange rate than the others, thus its rate determines the overall rate. Almost all the driving force drops over this step, the other steps are in quasi-equilibrium.*

Table 1 shows an exemplary reaction pathway, and the rate expressions for each of these steps being the rds. The reaction rate $\Re$ is the difference of forward and backward reaction rates $\vec{\Re}$,$\overleftarrow{\Re}$. Each of these rates can be expressed via a rate constant $\vec{k}$, $\overleftarrow{k}$ (which depends only on T) and the concentrations of species involved in the elementary step. As an example let us consider the dissociation step S3 to be rate determining, then

$$\Re = \vec{k}_3 \cdot [O_2^{2-}] - \overleftarrow{k}_3 [O^-]^2 \qquad (23)$$

For the backward reaction, $[O^-]$ appears squared because two O^- are reacting with each other. The exponents of the concentrations of reacting species in such rate expressions of elementary steps are called "reaction order", they typically have values of one (one molecule of this species involved, either unimolecular reaction as in the forward reactions of S1-S3, or reacting with another species as in S4) or two (two identical molecules/atoms reacting with each other, cf. backward reaction of S3). The reaction order for O_2 can furthermore take the value of 1/2 when the dissociation of molecular species is a fast preceding equilibrium (cf. S4). The determination of the reaction orders - in particular for O_2 - is very important because it allows one to draw strong conclusions on the reaction mechanism (order = 1 → molecular oxygen species in the rds; order = 1/2 → only atomic species in the rds and dissociation is a fast preceding equilibrium).

Table 1: *Elementary steps for an exemplary reaction pathway of oxygen incorporation. The mass action constants in the second column are valid when all reactions up to and including the respective step are fast (the concentration of regular oxide ions O_O^x is ≈ constant and included in K whenever it appears). Similarly, mass action constants for the subsequent fast reactions can be given. The rate equation in the last column is derived for the case that the respective step is the rds. The overall pO_2 dependence of $\Re_0$ is calculated for the hypothetical case of $[V_O^{\bullet\bullet}] \propto (pO_2)^{-0.1}$ and $[h^\bullet] \propto (pO_2)^{0.2}$.*

elementary step	mass action constants of preceding, subsequent react.	rate equation overall pO_2 dependence
S1: $O_{2,gas} \rightleftharpoons O_{2,ads}^- + h^\bullet$	$K_1' = \frac{[h^\bullet]^3}{[O_2^-][V_O^{\bullet\bullet}]^2}$	$\Re = \bar{k}_1 \cdot pO_2 - \bar{k}_1[O_2^-][h^\bullet]$ $= \bar{k}_1 \cdot pO_2 - \bar{k}_1 \frac{[h^\bullet]^4}{K_1'[V_O^{\bullet\bullet}]^2}$ (24) $\Re_0 \propto (pO_2)^1$
S2: $O_{2,ads}^- \rightleftharpoons O_{2,ads}^{2-} + h^\bullet$	$K_2 = \frac{[O_2^-][h^\bullet]}{pO_2}$ $K_2' = \frac{[h^\bullet]^2}{[O_2^{2-}][V_O^{\bullet\bullet}]^2}$	$\Re = \bar{k}_2 \cdot [O_2^-] - \bar{k}_2[O_2^{2-}][h^\bullet]$ $= \bar{k}_2 K_2 \frac{pO_2}{[h^\bullet]} - \bar{k}_2 \frac{[h^\bullet]^3}{K_2'[V_O^{\bullet\bullet}]^2}$ (25) $\Re_0 \propto (pO_2)^{0.8}$
S3: $O_{2,ads}^{2-} \rightleftharpoons 2\,O_{ads}^-$	$K_3 = \frac{[O_2^{2-}][h^\bullet]^2}{pO_2}$ $K_3' = \frac{[h^\bullet]}{[O^-][V_O^{\bullet\bullet}]}$	$\Re = \bar{k}_3 \cdot [O_2^{2-}] - \bar{k}_3[O^-]^2$ $= \bar{k}_3 K_3 \frac{pO_2}{[h^\bullet]^2} - \bar{k}_3 \frac{[h^\bullet]^2}{K_3'^2[V_O^{\bullet\bullet}]^2}$ (26) $\Re_0 \propto (pO_2)^{0.6}$
S4: $O_{ads}^- + V_O^{\bullet\bullet} \rightleftharpoons O_O^x + h^\bullet$	$K_4 = \frac{[O^-][h^\bullet]}{\sqrt{pO_2}}$	$\Re = \bar{k}_4 \cdot [O^-][V_O^{\bullet\bullet}] - \bar{k}_4[O_O^x][h^\bullet]$ $= \bar{k}_4 K_4 \frac{\sqrt{pO_2}}{[h^\bullet]}[V_O^{\bullet\bullet}] - \bar{k}_4'[h^\bullet]$ (27) $\Re_0 \propto (pO_2)^{0.2}$

The concentrations of intermediate species O_2^{2-} and O^- can be expressed via the mass action constants of fast preceding and subsequent equilibria, leading to the final rate expression as function of pO_2 and defect concentrations:

$$[O_2^{2-}] = \frac{K_2 pO_2}{[h^\bullet]} \qquad [O^-] = \frac{[h^\bullet]}{K_2'[V_O^{\bullet\bullet}]} \tag{28}$$

$$\Re = \bar{k}_3 K_3 \frac{pO_2}{[h^\bullet]^2} - \bar{k}_3 \frac{[h^\bullet]^2}{K_3'^2[V_O^{\bullet\bullet}]^2} \tag{29}$$

The other results of Table 1 are derived accordingly. In deriving these expressions, we assume that the coverage with adsorbed species is low, i.e. that the concentration of free adsorption sites is essentially constant (if this would be violated, an additional term for the concentration of free adsorption sites would appear). It is also possible that not the reaction step itself such as the incorporation $O^- + V_O^{\bullet\bullet} \rightleftharpoons O_O^x + h^\bullet$ is the bottleneck, but rather the necessary encounter of the reaction partners O^- and $V_O^{\bullet\bullet}$. Then, the mobility (or equivalently defect diffusivity) of

that reaction partner which migrates towards the other additionally appears in the rate equation (an experimental example is given in [20]).

The equilibrium exchange rate $\Re_0$ is given by the geometric mean of $\vec{\Re}$ and $\overleftarrow{\Re}$, and close to equilibrium (e.g. impedance measurement at open circuit) directly equal to $\vec{\Re}, \overleftarrow{\Re}$:

$$\Re_0 = \sqrt{\vec{\Re}\overleftarrow{\Re}}; \qquad \text{close to equil.: } \Re_0 = \vec{\Re} = \overleftarrow{\Re} \tag{30}$$

The effective rate constants k^δ, k^* and k^q are then related by eqs. (19-21) directly to $\Re_0$. The concentration c_O of regular oxide ions is constant, and also the thermodynamic factor w_O often depends only weakly on $p\mathrm{O_2}$, thus the $p\mathrm{O_2}$ dependence of the effective rate constants is largely given by that of $\Re_0$ (this $p\mathrm{O_2}$ dependence and dependence on defect concentrations is a big difference between the effective rate constants k^δ, k^*, k^q and the "elementary step rate constants" $\vec{k}, \overleftarrow{k}$ which depend only on T). It is important to note that the overall $p\mathrm{O_2}$ dependence of $\Re_0$ comprises also the contributions of the $p\mathrm{O_2}$ dependent defect concentrations. This is exemplarily calculated also in Table 1 for a hypothetical defect model with $[\mathrm{V_O^{\bullet\bullet}}] \propto (p\mathrm{O_2})^{-0.1}$ and $[\mathrm{h^\bullet}] \propto (p\mathrm{O_2})^{0.2}$. E.g., the $p\mathrm{O_2}$ dependence of the electronic defects leads to an overall $p\mathrm{O_2}$ dependence as low as 0.6 for S3 being the rds, although the reaction order for $\mathrm{O_2}$ is one. Since typically the defect concentrations decrease the overall $p\mathrm{O_2}$ dependence (examples in Table 1: $[\mathrm{V_O^{\bullet\bullet}}]$ being in the nominator and $[\mathrm{h^\bullet}]$ in the denominator for $\vec{\Re}$) an overall $p\mathrm{O_2}$ dependence larger than 0.5 means a reaction order for $\mathrm{O_2}$ of one (but vice versa a value below 0.5 does not necessarily prove an $\mathrm{O_2}$ reaction order of 1/2). While often it is more intuitive for the forward reaction, owing to the principle of microscopic reversibility, the $p\mathrm{O_2}$ dependence of the backward reaction rate must be identical.

One challenge in the interpretation of measured k values is that strictly speaking the defect concentrations appearing in the rate equations (and their $p\mathrm{O_2}$ dependences) are that of surface defects. In general the absolute values of surface defect concentrations are expected to differ from bulk values, and apart from few studies under realistic $p\mathrm{O_2}$, T conditions (e.g. [21]), experimental data are scarce. However, for the strongly acceptor-doped and redox-active perovskites used as SOFC cathode materials, often is it reasonable to assume the same majority defects for the surface as for bulk, which leads to similar $p\mathrm{O_2}$ dependences.

As long as they do not appear in the rds but only in preceding/subsequent equilibria, it does not matter if electronic defects are expressed as $\mathrm{e'}$ or $\mathrm{h^\bullet}$ because they are related via the band gap reaction $nil \rightleftharpoons \mathrm{e'} + \mathrm{h^\bullet}$. Only if they appear in the rds, this would lead to different overall $p\mathrm{O_2}$ dependences. Since electronic carriers are involved in preceding fast equilibria such as step S1, a correlation between electronic structure parameters (such as the band gap in [22]) does not necessarily prove that electron transfers are involved in the rds; still the dissociation or an ion transfer step (S4) may be rate determining.

For the elucidation of the reaction mechanism, ab initio calculations are an important complementary tool yielding information that is very difficult to obtain from experiments (energies of surface defects and intermediate species, reaction or surface migration barriers). For mixed conducting oxides with a variable and perceptible nonstoichiometry, already the measurement of adsorbate concentrations is very difficult because desorption from the surface and excorporation from the bulk strongly overlap. Adsorption energies of molecular ($\mathrm{O_2^-}$, $\mathrm{O_2^{2-}}$) and atomic species ($\mathrm{O^-}$) are negative, but the large negative $\Delta_{ads}S^0$ leads to low adsorbate coverage at typical SOFC operation temperatures (see e.g. [23,24]).

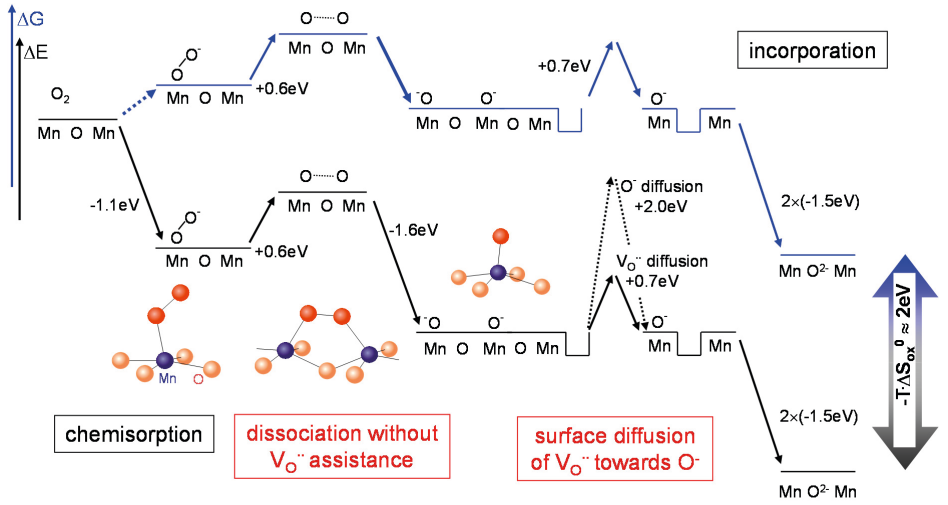


Fig. 5: *Energy (black) and Gibbs free energy profile (blue) for oxygen incorporation into* $V_O^{\bullet\bullet}$ *at the (001)MnO_2 termination of $LaMnO_3$ obtained from DFT calculations. For the $\Delta_r G$ profile, the entropic contributions at a typical SOFC operation temperature of 1000 K are estimated on the basis of the overall reaction entropy of ≈ -200 J/mol K (i.e. losing almost all the entropy of the gas-phase O_2; the largest part occurring in the adsorption step). For the encounter of* $V_O^{\bullet\bullet}$ *and* O^-*, one finds a much lower surface migration barrier of 0.7 eV for* $V_O^{\bullet\bullet}$ *compared to 2 eV for* O^-*, thus* $V_O^{\bullet\bullet}$ *approaches* O^-*. The final incorporation of neighboring* O^- *and* $V_O^{\bullet\bullet}$ *has no perceptible barrier. Adapted from* [24].

Another important result from ab initio calculations is that the dissociation of molecular oxygen species such as adsorbed peroxide O_2^{2-} on a perfect perovskite surface has a perceptible barrier (e.g. 0.6 eV for $LaMnO_3$, (001)MnO_2 termination [24], Fig. 5). On the other hand, with assistance by a $V_O^{\bullet\bullet}$ the barrier becomes negligible (peroxide first vertically dropping into the vacancy, then splitting the O-O bond leading to one adsorbed O^- and one O_O^x filling the $V_O^{\bullet\bullet}$). However, for (La,Sr)MnO_3, the $V_O^{\bullet\bullet}$ concentration is so low that the dissociation without $V_O^{\bullet\bullet}$ assistance is the faster option. In Fig. 5 the $\Delta_r G$ profile exhibits two maxima of comparable height (dissociation, and mutual approach of $V_O^{\bullet\bullet}$ and O^-, indicated in red) which are the potential rate determining steps. It then depends on the detailed experimental conditions (e.g. pO_2) which of the two actually becomes limiting.

Altogether, the oxygen exchange reaction requires ionic defects ($V_O^{\bullet\bullet}$) as well as the transfer of electrons to adsorbed oxygen species. As a rough guideline, one can state that for an "electron-poor" material with large band gap and small redox activity such as slightly acceptor-doped $SrTiO_3$, the steps involving electron transfer (or directly subsequent such as O_2^{2-} dissociation) tend to be limiting [25]. On the other hand, for "electron-rich" materials with small band gap and high redox activity, rather the steps related to oxygen vacancies become the bottleneck [26,20], and generally the exchange rates are significantly higher than for "electron-poor" materials.

2.4 Kinetics of oxide scale formation

During the redox processes discussed in sections 2.2 and 2.3 the material remained single-phase. In contrast, the process of scale (oxide layer) formation on metals discussed in the present section implies formation of an additional solid phase, cf. reaction (2). Scale formation has high technical relevance; the development of a dense surface oxide layer preventing further corrosion is decisive for many applications of non-noble metals. Two criteria determine the kinetics of scale formation: (i) In order to obtain a well-adhering oxide layer, the molar volumes of metal and oxide have to be quite similar, otherwise crack formation or flaking of the oxide repeatedly expose fresh metal. (ii) The kinetics of materials transport through the dense oxide layer must be sufficiently slow.

Ref. [7] gives a good overview of different regimes of oxide scale formation. The main distinction is whether electric fields developing across the film are important ("Cabrera Mott theory" for thin films) or not ("Wagner theory" for thick films). Further complications by space charge effects within the growing oxide film are not discussed here - please refer to [7,27].

Wagner theory [28,7]

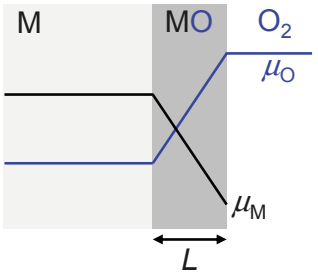


Fig. 6: *Growth of a dense oxide film of thickness L on a metal when no relevant electric fields are present across the film. The slope of μ_O and μ_M within the film is not necessarily linear (deviations occur when the relevant D* is not constant but varies with μ_O, μ_M).*

The situation for "thick" films is depicted in Fig. 6. The driving force for oxide formation is the difference in μ_O between the metal and the gas atmosphere. The metal has a nonzero oxygen solubility; μ_O in the metal is defined by the coexistence with the oxide at given T, see eq. (7). The gradient in μ_O - and correspondingly opposite gradient in μ_M (Gibbs Duhem relation) - leads to in-diffusion of oxygen and/or out-diffusion of metal, depending on the relative magnitude of metal and oxygen defect concentrations and mobilities. M and O diffusion are both ambipolar diffusion processes of ionic and electronic defects. Since $\nabla\mu_O$ decreases with increasing film thickness (the same overall $\Delta\mu_O$ drops over a larger L), the growth rate decreases with time resulting in a parabolic rate law. The derivation is formulated here in terms of metal diffusion (in most binary oxides, cations are more mobile than oxide ions). The flux of cations must be balanced by a corresponding flux of electronic carriers, leading to

$$j_{M^{2+}} = j_M = -\frac{\sigma_{M_i^{\bullet\bullet}}\sigma_{e'}}{4F^2(\sigma_{M_i^{\bullet\bullet}}+\sigma_{e'})}\frac{\partial\mu_M}{\partial x} \qquad (31)$$

for the case that the ionic carriers are metal interstitials and the electronic carriers excess electrons (cf. expression for chemical diffusion coefficient in chapter A4). This expression emphasizes that a perceptible ionic as well as electronic conductivity is required for high rates of

scale formation. The metal flux is constant within the oxide film (no sink terms inside the MO layer), and integration of eq. (31) leads to

$$j_M = -\frac{1}{4F^2 L}\int_{\mu_M(x=0)}^{\mu_M(x=L)} \frac{\sigma_{M_i^{\cdot\cdot}}\sigma_{e'}}{\sigma_{M_i^{\cdot\cdot}}+\sigma_{e'}} d\mu_M \qquad (32)$$

Since $j_M = V_{MO} \cdot dL/dt$ with V_{MO} = molar volume of MO, this is equivalent to

$$\frac{dL}{dt} = \frac{\kappa}{L} \qquad \Rightarrow \qquad L^2 = 2\kappa t \qquad (33)$$

with the "parabolic rate constant" κ given by

$$\kappa = -\frac{V_{MO}}{4F^2}\int_{\mu_M(x=0)}^{\mu_M(x=L)} \frac{\sigma_{M_i^{\cdot\cdot}}\sigma_{e'}}{\sigma_{M_i^{\cdot\cdot}}+\sigma_{e'}} d\mu_M = -\frac{V_{MO}}{4F^2}\int_{\mu_M(x=0)}^{\mu_M(x=L)} \sigma_{M_i^{\cdot\cdot}} t_{eon} d\mu_M \qquad (34)$$

where $t_{eon} = \sigma_{eon}/(\sigma_{eon}+\sigma_{ion})$ is the electronic transference number (close to unity for binary oxides such as ZnO, CoO, NiO). While in principle both $\sigma_{M_i^{\cdot\cdot}}$ and t_{eon} vary with μ_O across the oxide layer, they can be represented by averaged values (indicated by $\langle\ \rangle$; for explicit consideration of this μ_O dependence see e.g. [1]), and the integration of $d\mu_M$ yields $\Delta\mu_M$:

$$\kappa = -\frac{V_{MO}\left\langle \sigma_{M_i^{\cdot\cdot}} t_{eon}\right\rangle}{4F^2}\Delta\mu_M = -\left\langle D_M^* t_{eon}\right\rangle \frac{\Delta G_{MO}^0}{RT} \qquad (35)$$

The Nernst-Einstein relation was used to substitute σ_{ion} by the self (≈ tracer) diffusivity of the mobile ionic defect. ΔG_{MO}^0 is the standard formation Gibbs energy of the oxide MO from M and O_2. Eq. (35) clearly shows that the "parabolic rate constant" is not a surface rate constant (as the k's in section 2.2, 2.3) but instead closely related to diffusivities. It also demonstrates another peculiarity of scale growth: while the nature of the diffusion process is chemical diffusion (coupled transport of ionic and electronic carriers) the resulting time dependence eq. (33) differs from that of stoichiometry relaxation of a single-phase sample (eqs. (13-15)). This is related to the fact that in scale growth the driving force remains constant ($\Delta\mu_O$, $\Delta\mu_M$) and the the resistive contribution grows proportionally to the film thickness L, while in stoichiometry relaxation the driving force decays.

In many binary oxides, the bulk ionic diffusivity is very low (much lower than oxygen diffusivity in acceptor-doped perovskites and fluorites), thus extended defects such as dislocations and grain boundaries are often found to lead to enhanced diffusion, and to dominate the scale growth kinetics (see e.g. [7]). The low diffusivities also mean that the critical thickness $l = D/k$ below which the surface reaction would determine the kinetics becomes very small. Typically, such a regime is not unambiguously observed, and for very thin films rather the Cabrera Mott theory applies.

Cabrera Mott theory [29,7]

The scale formation theory by Cabrera and Mott describes the regime of extremely thin films where a very strong electric field across the film significantly affects the ion transport. The field arises from the fact that typically the work function of the metal is smaller than the ionization potential of negatively charged chemisorbed oxygen species at the surface of the oxide. Consequently, electrons will be transferred from the metal through the initially very thin oxide film (by tunneling or thermionic emission) to form negative oxygen adsorbates on the oxide surface, until the electrochemical potential of the electrons is equilibrated throughout

the system (Fig. 7). The field $\Delta\phi/L$ is the larger the thinner the oxide film, and can reach a magnitude of 10^9 V/m ($\Delta\phi$ of 1-2 V, oxide thickness of few nm).

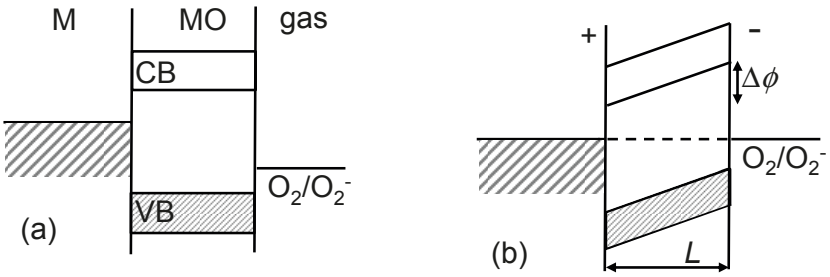


Fig. 7: *Situation for a very thin oxide film MO on a metal M (a) before equilibration of the electrochemical potential of the electrons, (b) after equilibration by electron transfer from the metal to chemisorbed oxygen species at the film surface.*

In the Cabrera Mott theory it is assumed that owing to the extremely small oxide thickness, ion transfer at the inner interface or at the surface is rate limiting (not carrier transport within the film). The respective ion transfer barrier is partly decreased by the electric field (analogously to the Butler Volmer approach for electrode kinetics). This leads to a field-dependent term appearing in the rate equation (e_0 = elementary charge, a = jump distance, k = Boltzmann constant):

$$\frac{dL}{dt} \propto \exp\left(\frac{e_0 a \Delta\phi}{2kTL}\right) \qquad \Rightarrow \qquad \frac{1}{L}\frac{e_0 a \Delta\phi}{2kT} \approx const. - \ln(t) \tag{36}$$

The oxidation rate decreases exponentially with increasing L, corresponding to an inverse logarithmic equation for the film thickness. The strong decrease of the growth rate with increasing L leads to a "limiting thickness" in the range of 10 nm (cf. [7]) above which this growth mode ceases. Only if the temperature is higher than a critical value (which is related to the barrier height of the limiting process, cf. [7]), the growth then continues in the parabolic regime.

3 Redox processes in electrochemical cells

3.1 Electrochemical cells at open circuit

Electrochemical cells consist of a (predominantly) ion conducting material that separates the two electrodes at which redox processes occur. The reaction partners may be solid, liquid or gaseous. The fact that the electrolyte membrane separates the space of reactants and products, and that it (predominantly) conducts only ions forces the chemical reaction under investigation to proceed via separate ion transfer (through the electrolyte) and electron transfer (through the outer circuit, where the electron flow can perform work). For simplicity let us consider the "reaction" of transferring oxygen from high $p\mathrm{O_2}$" to low $p\mathrm{O_2}$ ' (Fig. 8). "Open

circuit" means that no current is flowing in the outer electrical circuit. For electrolytes with $t_{eon} = 0$ also no current flows within the cell (Fig. 8a). However, for electrolytes with a small but nonnegligible t_{eon}, an ionic and electronic current flow within the cell in opposite direction (Fig. 8b). This corresponds to some permeation of a neutral component through the electrolyte membrane (e.g. oxygen permeation by coupled transport of $V_O^{\bullet\bullet}$ and e').

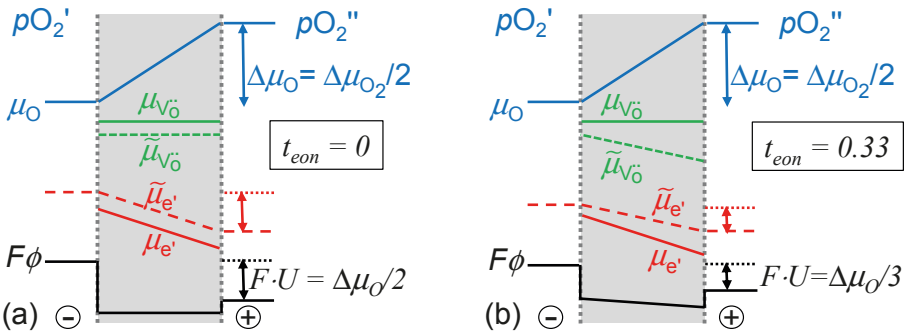


Fig. 8: *Electrochemical cells with $pO_2' < pO_2''$ at open circuit. (a) electrolyte with $t_{eon} = 0$ (e.g. YSZ) ⇒ no internal current (b) electrolyte with $t_{eon} \neq 0$ (e.g. Gd-doped CeO_2 at $T > 600$ °C; the value of $t_{eon} = 0.33$ is a bit overemphasized to show the effects more clearly) which leads to internal ionic and electronic currents. For both cases the electrode reactions are assumed to be sufficiently fast (negligible electrode overpotential) so that the measured voltage is entirely determined by equilibrium properties.*

Without being separated by an electrolyte membrane, this "reaction" simply corresponds to the expansion of O_2, with $\Delta G = -RT\ln(pO_2''/pO_2')$. Chemical equilibrium is achieved when $\Delta G = 0$, i.e. $pO_2'' = pO_2'$, and such a system in chemical equilibrium cannot deliver any work to the environment. When the two gas spaces are separated by an oxide ion conducting membrane (mobile O_i'' or $V_O^{\bullet\bullet}$), the reaction can proceed only by splitting the oxygen into ionic and electronic carriers at both electrodes:

$$1/2\ O_2(pO_2') + 2\ e^- \text{ (left side)} \rightleftharpoons O^{2-}\text{(left side)} \tag{37}$$

$$1/2\ O_2(pO_2'') + 2\ e^- \text{ (right side)} \rightleftharpoons O^{2-}\text{(right side)} \tag{38}$$

The high ionic conductivity of the electrolyte forces the electrochemical potential of the oxide ions to be equal in the electrolyte and at both electrodes. Since μ_O differs between both sides, this leads to the difference in electrochemical potential of the electrons that is measured as voltage in the outer circuit. The cell is not in chemical equilibrium, and for the $t_{eon} = 0$ case ΔG of the cell (including the actual concentrations/activities of the reaction partners, thus possibly deviating from ΔG^0) is completely converted into electrical energy

$$\Delta G = -zFU \tag{39}$$

where z is the number of electrons transferred in the reaction. Let us discuss the (electro)chemical profiles in the cells in more detail. The electrochemical potential $\tilde{\mu}_i$ of a species i is composed of its chemical potential μ_i its charge z_i and the electrical potential

$$\tilde{\mu}_i = \mu_i + z_i F\phi \tag{40}$$

The measured voltage (electromotive force, emf) is related to the difference in electrochemical potential of the electrons. For the $t_{eon} = 0$ case we obtain the Nernst equation

$$\begin{aligned} U \cdot F &= -\Delta\tilde{\mu}_{e'} = -\frac{1}{2}\left(\Delta\tilde{\mu}_{O^{2-}} - \Delta\mu_O\right) = -\frac{1}{2}\left(-\Delta\tilde{\mu}_{V_O^{\cdot\cdot}} - \Delta\mu_O\right) \\ &= \frac{1}{2}\Delta\mu_O = \frac{1}{4}RT\ln\frac{pO_2''}{pO_2'} \end{aligned} \tag{41}$$

using the relations $\tilde{\mu}_{e'} = \tilde{\mu}_{O^{2-}} - \mu_O$, $\tilde{\mu}_{O^{2-}} = -\tilde{\mu}_{V_O^{\cdot\cdot}}$ and $\mu_O = \mu_{O_2}/2$. $\mu_{V_O^{\cdot\cdot}}$ and $\tilde{\mu}_{V_O^{\cdot\cdot}}$ have flat profiles throughout the electrolyte because the $V_O^{\cdot\cdot}$ concentration is constant (fixed by the dopant in YSZ) and the interior is field-free (highly ion-conducting material but $j = 0$ because of open circuit conditions). While there is no electrical potential gradient within the YSZ, potential steps (corresponding to double layers at hetero-interfaces) are present at the interfaces to the electrodes. Formally, one has to consider a minute electronic conductivity in the YSZ in order to have well-defined $\tilde{\mu}_{e'}, \mu_{e'}, \mu_O$ within the electrolyte.

For the case with $t_{eon} \neq 0$ shown in Fig. 8b, the situation differs in some aspects. The ionic and electronic current compensate each other

$$j_{V_O^{\cdot\cdot}} = -\frac{\sigma_{V_O^{\cdot\cdot}}}{4F^2}\nabla\tilde{\mu}_{V_O^{\cdot\cdot}} \qquad j_{e'} = -\frac{\sigma_{e'}}{4F^2}\nabla\tilde{\mu}_{e'} \quad \rightarrow \quad 2\sigma_{e'}\nabla\tilde{\mu}_{e'} = -\sigma_{V_O^{\cdot\cdot}}\nabla\tilde{\mu}_{V_O^{\cdot\cdot}} = \sigma_{O^{2-}}\nabla\tilde{\mu}_{O^{2-}} \tag{42}$$

$\tilde{\mu}_{e'}$ is obtained by considering also the local equilibrium

$$\mu_O = \tilde{\mu}_{O^{2-}} - 2\tilde{\mu}_{e'} \quad \rightarrow \quad \nabla\tilde{\mu}_{e'} = \frac{\sigma_{O^{2-}}}{\sigma_{O^{2-}} + \sigma_{e'}}\nabla\mu_{O_2} \tag{43}$$

$$U = \frac{1}{4F}\int_{\mu_{O_2}'}^{\mu_{O_2}''} t_{ion}\, d\mu_{O_2} = \frac{RT}{4F}\langle t_{ion}\rangle \ln\frac{pO_2''}{pO_2'} = -zF\langle t_{ion}\rangle\Delta G \tag{44}$$

and the voltage is obtained after integration over $\nabla\tilde{\mu}_{e'}$. $\langle t_{ion}\rangle$ is the ionic transference number averaged over the covered pO_2 range, and it determines how much the measured voltage is lower that the Nernst voltage.

Regarding the profiles within the electrolyte, $\mu_{V_O^{\cdot\cdot}} = const$ is still a reasonable assumption because the changes in $V_O^{\bullet\bullet}$ concentration caused by the slight redistribution of electronic defects in the potential gradient is small in relative terms. Because of the local equilibrium with μ_O, also $\mu_{e'}$ has the same slope as in Fig. 8a. The slope of $\tilde{\mu}_{e'}$ is decreased according to the lower voltage in eq. (44), and the difference yields the electrical potential gradient $\nabla\phi$ in the electrolyte (nonzero, in contrast to Fig. 8a). $\nabla\phi$ then also causes the gradient in $\tilde{\mu}_{V_O^{\cdot\cdot}}$ shown in Fig. 8b. By applying eq. (44), electrochemical cells at open circuit can be used to determine the ionic transference number of mixed conductors (as long as t_{ion} exceeds several percent).

An important technical application of cells at open circuit are potentiometric gas sensors such as the lambda probe to measure pO_2 in exhaust gas for adjusting the fuel/air ratio. The measured gas is not necessarily identical to the mobile ionic species, e.g. the reaction $Na_2CO_3 \rightleftharpoons CO_2 + 1/2\ O_2 + 2\ Na^+ + 2\ e^-$ at the electrode couples pCO_2 to the Na^+ activity, allowing to build a potentiometric CO_2 sensor based on a Na^+ conducting electrolyte. The combination with an air-stable two-phase $Na_2Ti_6O_{13}/TiO_2$ reference electrode leads to a sensor signal that is not disturbed by pO_2 changes [30].

In a fuel cell, the effective pO_2 at the anode is determined by the gas composition (e. g. pH_2 and pH_2O) and the respective equilibrium constant (for $H_2 + 0.5\ O_2 \rightleftharpoons H_2O$). The resulting

pO_2 for equal pH_2 and pH_2O amount to about 10^{-63} bar at 100 °C and 10^{-20} bar at 700 °C. Correspondingly, the open circuit voltage of the fuel cell (which can also directly be obtained from $zFU = -\Delta G$) decreases from 1.17 V to 1.01 V with increasing T.

3.2 Cells generating current or acting as electrochemical pump

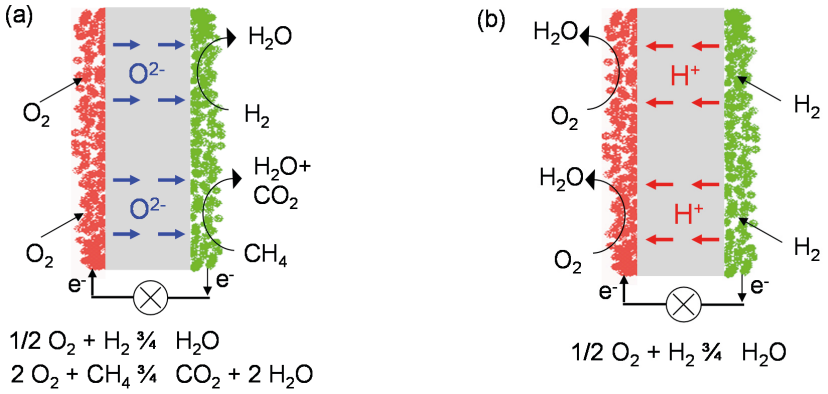


Fig. 9: *Fuel cells based on (a) oxide ion conducting electrolyte (e.g. Y-doped ZrO_2, Gd-doped CeO_2, $T \geq 600$ °C). (b) proton conducting electrolyte - either proton conducting polymer such as Nafion ($T \approx 80$ °C, see e.g.* [31] *and references therein) or proton conducting ceramic membrane such as Y-doped $BaZrO_3$ (T = 400-600 °C, see e.g.* [32] *and references therein). The overall reactions are indicated at the bottom.*

Electrochemical cells for electricity generation comprise batteries (primary batteries, or rechargeable) and fuel cells. While the former have only that amount of chemical energy available for conversion that is initially stored in the electrode materials, the liquid or gaseous fuel (anode side) and oxygen (air; cathode side side) in fuel cells can be replenished continuously. Both types of cells have the advantage that they are not limited by the Carnot efficiency, but rather by eq. (44). As soon as a current is drawn from electrochemical cells, the voltage drops below the open circuit voltage. The higher the current density, the higher the voltage drop (losses) and the lower the conversion efficiency.

Fig. 9 illustrates the main types of fuel cells: (i) based on oxide ion conducting electrolytes such as YSZ. They typically require operation temperatures above 600 °C, but since it is oxide ions that are transported through the ceramic electrolyte membrane, they can operate on hydrogen as well as on hydrocarbon fuels (with some measured taken to prevent coke deposition on the anode). (ii) based on proton conducting electrolytes. There are two options: proton conducting polymers operating at ≈ 80 °C, and proton conducting ceramic membranes operating at 400-600 °C.

Several processes contribute to the losses as exemplified in Fig. 10, based on numerical values that are in a range found for SOFC. The ohmic resistance (mainly from the electrolyte) leads to a linear voltage drop according to $\Delta U = R_{ohm} \cdot I$. The electrode reactions are in general not fully reversible (not infinitely fast), and thus also lead to a voltage drop. In contrast to the

ohmic loss, the electrode reaction resistance is not constant but decreases with increasing current density (more details on electrode kinetics in the next section). Thus, the loss from the reaction resistance is most severe at low current density and then levels off (Fig. 10a). On the other hand, at high current densities, gas phase transport limitations and/or pore diffusion come into play. The overall result is that the power density passes through a maximum as shown in Fig 10b. The conversion efficiency drops e.g. to only 56 % at 1.2 A/cm^2.

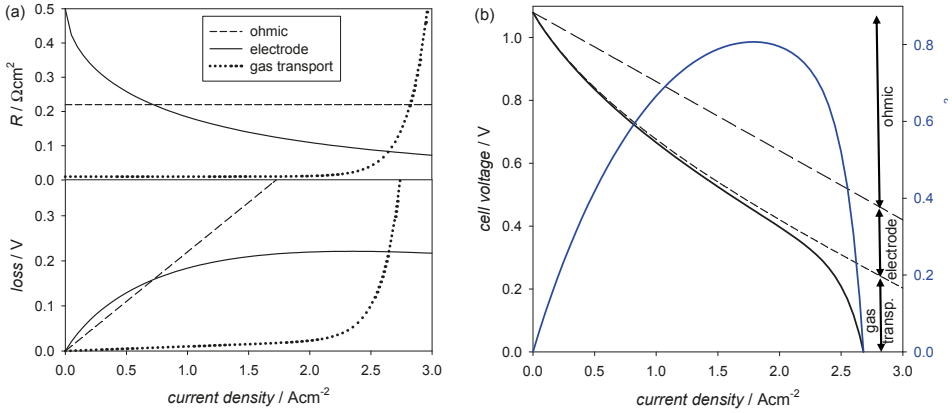


Fig. 10: *(a) Resistances, losses (b) cell voltage and power density for a SOFC. The relative magnitude of the different losses varies with the current density.*

Instead of generating electricity, an electrochemical cell can also be driven by an externally supplied current. When an oxide ion conducting electrolyte is used, such a device can be applied to pump oxygen from one side to the other with an extremely high selectivity and exactly controlled oxygen flux, e.g. for supplying oxygen into a catalytic reaction. When a fuel cell is driven in backward direction by an external current, it works as an electrolyzer, inverting the reactions given in Fig. 9 converting e.g. water to hydrogen (see e.g. [33]). High temperature electrolysis cells benefit from the voltage decrease calculated above for the reaction $H_2O \rightleftharpoons H_2 + 1/2\ O_2$ cell with increasing T. In electrolysis mode, heat created by internal resistances as soon as a current flows can contribute to driving the water splitting reaction.

3.3 Electrode kinetics

For cells at open circuit, the electrode kinetics is determined by the equilibrium exchange rate as discussed in section 2.2. Under current flow, the corresponding potential gradients modify the reaction rate. For electrodes in contact with liquid electrolytes, this dependence can often be described by the Butler Volmer equation (see e.g. [34] for detailed derivation)

$$i = i_0 \left[\exp\left(\bar{\alpha} \frac{zF\eta}{RT} \right) - \exp\left(-\bar{\alpha} \frac{zF\eta}{RT} \right) \right] \tag{45}$$

with i_0 = exchange current density, η = overpotential, $\bar{\alpha}, \bar{\alpha}$ = symmetry factors ($\bar{\alpha} + \bar{\alpha} = 1$). The simplistic picture behind this equation is that a part of the applied overpotential ($\bar{\alpha}\eta$ for forward and $\bar{\alpha}\eta$ for backward reaction) lowers the effective reaction barrier. While for elec-

trodes in electrolyte solution η indeed drops at the electrode surface, the surface potential drop of gas-solid electrodes - which is the quantity that affects the electrode kinetics - is not necessarily identical to η. The potential drop at the surface may well change with applied η, but it is also influenced by the coverage with charged adsorbates (which itself may depend on η). An important difference to metal electrodes is that in electrodes consisting of mixed conducting oxides an applied η corresponds to a change of μ_O, which may change the concentrations of defects that are relevant for the reaction rate. For more details see [35,36]. While for a certain overpotential range an expression of the form of eq. (45) - maybe with $\bar{\alpha}+\bar{\alpha}\neq 1$ - could be able to fit measured data, care must be taken regarding a physically correct interpretation.

3.4 Stoichiometry polarization, Wagner-Hebb experiments

The term stoichiometry polarization describes an experiment in which a current is drawn through a sample having more than one mobile carrier, and where at least one electrode is selectively blocking. The purpose of this experiment is to determine the conductivity of the non-blocked carrier (see [37,38,39]). From the transient behavior also the chemical diffusion coefficient can be obtained.

Fig. 11a gives an experimental example: galvanostatic measurements on mixed conducting $Ag_{1.93}Te$ with ion- as well as electron-blocking electrodes. Directly after the switching on of the current, the voltage drop between the sensing probes represents the resistance from the total conductivity $U_{t=0} = R_{tot} \cdot I$, because also the carrier for which the electrode is not permeable can migrate some distance within the sample before "hitting" the selectively blocking interface. The conductivity calculated from this short-time response is identical for ion- and electron-blocking electrodes. At long time, the current of the blocked carrier is zero, i.e. the steady state current is carried only by the non-blocked carrier and the voltage drop is related to the non-blocked conductivity by $U = R_{non\text{-}blocked} \cdot I$. After switching off the current, the relaxation of U occurs with the same transient behavior as for the polarization. In Fig. 11a the experiment is shown for both ion- and electron blocking (consistently yielding $t_{eon} \approx 0.2$); typically the measurement is performed only for the carrier with the smaller transference number. The comparison of the DC measurement to AC (impedance response) of a sample contacted with one or two selectively blocking electrode can be found in [1]; both techniques are in principle able to resolve σ_{ion} or σ_{eon} and σ_{tot}, but when the approach to the steady state takes very long the time-resolved DC measurement is more practical than the frequency-resolved impedance spectroscopy.

Fig. 11b schematically shows the concentration profiles that develop with time in the sample. The blocked carrier first accumulates only close to the blocking interface, and with time this concentration gradient extends over the whole sample thickness. It balances the electrical driving force felt by this carrier, finally leading to a complete ceasing of its current. Owing to local electroneutrality, also the mobile carrier develops a corresponding gradient. Thus, the formed gradient is the concentration gradient of a neutral component. Therefore it is also clear that the transient behavior of the voltage drop is determined by chemical diffusion of this neutral component, following a $\sqrt{t/\tau^{\delta}}$ behavior for short and an exponential decay for long time (similar to eqs. (13-14)). The characteristic time τ^{δ} is identical for both experiments in Fig. 11a, and related by

$$\tau^{\delta} = \frac{4L^2}{\pi^2 D^{\delta}} \tag{46}$$

to the chemical diffusion coefficient (L = full sample thickness; when both electrodes are blocking the factor 4 has to be omitted).

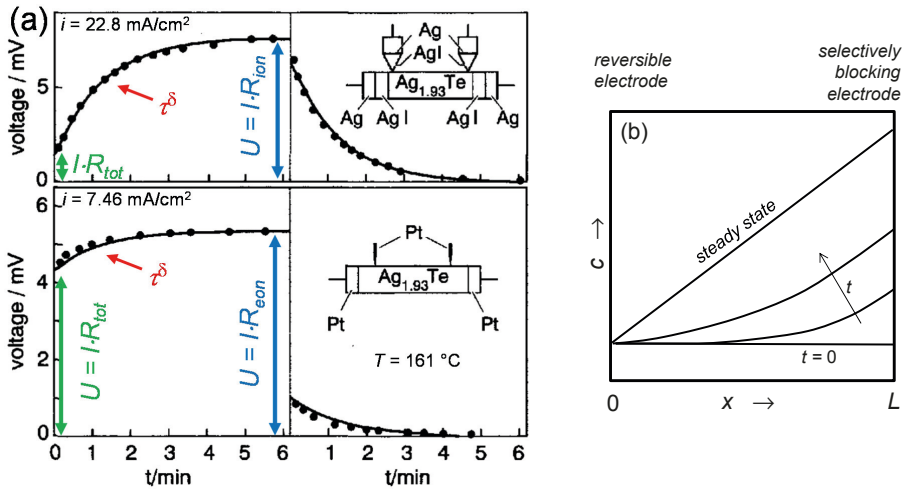


Fig. 11: *(a) Galvanostatic measurement on $Ag_{1.93}Te$ with electron-blocking (top) and ion-blocking electrodes (bottom). The voltage sensing probes are separated from the current contacts to eliminate interfacial transfer resistances. $Ag_{1.93}Te$ has an ionic transference number of ≈ 0.2 under measurement conditions. Adapted from* [1,39]. *(b) Concentration profiles under galvanostatic polarization; the left electrode is reversible for ionic and electronic carriers, the right electrode selectively blocks one of the carriers. Adapted from* [40].

For practical experiments with oxides one has to ensure that a "leakage" through the lateral sides is avoided (gas-tight sealing; if T is low enough to freeze out the exchange kinetics of the bare surface this can be omitted but then the reversible electrode requires a good O exchange catalyst). Furthermore, voltage drops at the interface to the selectively blocking electrode should be avoided (or at least quantified by impedance spectroscopy), for some other critical points see e.g. [41].

So far the discussion referred to comparably small voltage drops created by the polarization, thus the extracted D^{δ} and t_{ion}, t_{eon} which depend more or less strongly on the exact sample nonstoichiometry, are averaged only over a small composition range. However, one can deliberately enforce large voltage drops corresponding to large gradients in component activity within the sample to actually extract the dependence of t_{ion}, t_{eon}, D^{δ} on this activity (i.e. on the effective pO_2 in an oxide sample). For quantitative details see e.g. [1].

4 Processes driven by voltage load or *T* gradients

In this section we deal with some more cases of ion transport driven by potential gradients (kinetic demixing, thermodiffusion) which are not covered in sections 2 and 3. They can become critical for the long-time stability of electrochemical devices. More details can be found e.g. in [8,9,42] which cover also related effects such as kinetic decomposition or morphological instabilities.

4.1 Kinetic demixing

The term kinetic demixing means the transformation of a homogeneous solid solution into an inhomogeneous system (but still single-phase) by a potential gradient. After a transient, the system will reach a state with steady concentration gradients. They start to revert as soon as the potential gradient is not active any more. The driving force can be an electrical potential gradient as well as a chemical potential gradient.

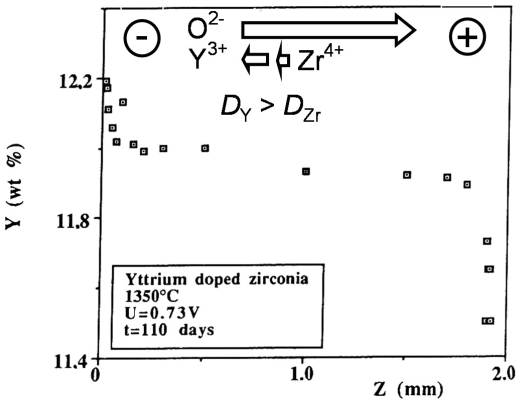


Fig. 12: *Kinetic demixing of Y and Zr in Y-doped ZrO_2 under voltage load of 0.73 V. Note that the Y concentration profile is still far from the steady state. Figure taken from* [43].

An example for the first case is Y-doped ZrO_2 (YSZ, an electrolyte conducting oxide ions via oxygen vacancies) under a DC current. Even for reversible electrodes (i.e. fast oxygen exchange reaction $1/2\ O_2 + V_O^{\bullet\bullet} + 2e' \rightleftharpoons O_O^x$), a small part of the current will be carried by cation migration (the concentration of cation defects and their mobility are much lower than that of $V_O^{\bullet\bullet}$ but not zero). The kinetic demixing is caused by the fact that the charge and mobility of Y^{3+} and Zr^{4+} differ (irrespective of the mechanism of cation migration proceeding via vacancies or interstitials). As a consequence, the more mobile Y^{3+} accumulates close to the cathode (Fig. 12). The magnitude of the steady state concentration changes increases with applied voltage. Such a demixing can become problematic when a YSZ electrolyte membrane is operated for very long times in a SOFC, because then in large parts of the membrane the Y content deviates from the the optimum dopant content.

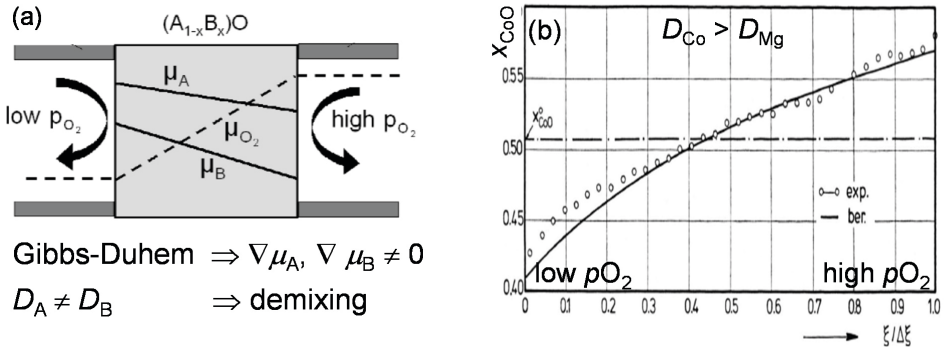


Fig. 13: *Kinetic demixing in a pO_2 gradient. (a) Sketch of chemical potential gradients, figure taken from* [9]. *(b) Experimental results for $Co_{0.6}Mg_{0.4}O$ exposed to a pO_2 gradient ($p'/p'' = 3$), the more mobile Co accumulates at the high pO_2 side. Figure taken from* [44].

An example for the second case (chemical potential gradient) is a $(Co_{1-x}Mg_x)O$ sample exposed to different pO_2 at the two sides. This causes not only a μ_O gradient in the sample, but according to the Gibbs-Duhem relation (eq. (47), x denotes molar fraction) also gradients in μ_{Co} and μ_{Mg} as shown in Fig. 13a. The relative magnitude of $\nabla\mu_{Co}$ and $\nabla\mu_{Mg}$ is determined by the ratio of the cation diffusivities:

$$x_{Co}d\mu_{Co} + x_{Mg}d\mu_{Mg} + x_O d\mu_O = 0 \tag{47}$$

$$D_{Co}\nabla\mu_{Co} = D_{Mg}\nabla\mu_{Mg} \tag{48}$$

The more mobile cation accumulates at the side with higher pO_2 (Fig. 13b).The larger the ΔpO_2, the larger the accumulation. This mode of kinetic demixing can become detrimental for oxygen permeation membranes.

4.2 Thermodiffusion

Temperature gradients can also act as the driving force for transport. The key quantity is the "heat of transport", i.e. the heat that is carried along when a carrier is transported. This couples electrical or mass flux and heat flux. The Seebeck effect of electronic carriers driven by ΔT is widely applied in the temperature measurement by thermocouples. Thermoelectric devices based on the Seebeck effect are intensively investigated for direct electricity generation from waste heat. Also the inverse process - Peltier effect, generation of a T gradient by and electric current - is well known and widely applied.

The respective effects of coupling heat and mass transfer are the Ludwig-Soret effect (thermodiffusion, ion transport driven by ΔT) and the Dufour effect (ΔT generated by ion flux). The Soret effect can lead to gradients in the nonstoichiometry of a binary material (e.g. $Cu_{2-\delta}O$ [42]) but also to gradients of a dopant in a host material (e.g. M^{2+} donor dopants in alkali halides, see refs. given in [45]), or to the demixing of solid solutions ($Co_{1-x}Mg_xO$ [46], Fig. 14). These processes can affect the long-term stability of devices operating in T gradients.

For mixed-conducting $Cu_{2-\delta}O$ [42], the electronic as well as the ionic thermopower was measured. The relative change of the copper nonstoichiometry δ under a temperature gradient was found to be 1.7 % per K. Thus, already moderate *T* gradients of the order of 60 K might lead to the decomposition of this material. The extracted heat of transport of 180 kJ/mol for Cu in $Cu_{2-\delta}O$ is of fundamental interest for the understanding of ion transport mechanisms, its magnitude can exceed the migration barrier.

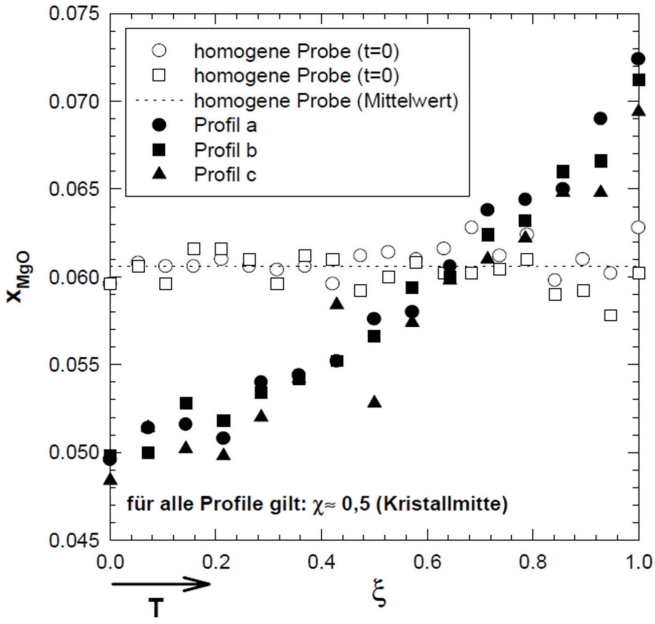


Fig. 14: *Demixing of a $Co_{0.6}Mg_{0.4}O$ single crystal after 214 h in a temperature gradient of 45 K at a mean T = 1528 K. Figure taken from* [46].

References

[1] J. Maier, *Physical Chemistry of Ionic Materials*, Wiley (2004).

[2] H. Mehrer, *Diffusion in Solids*, Springer (2007).

[3] H. Schmalzried, *Chemical Kinetics of Solids*, VCH (1995).

[4] J. Maier, Solid State Ionics 112 (1998) 197 and 135 (2000) 575.

[5] R. Merkle, J. Fleig, J. Maier, in: *Handbook of Fuel Cells*, ed. H. Yokokawa, Wiley (2009).

[6] R. Merkle, J. Maier, Angew. Chemie Int. Ed. 47 (2008) 2.

[7] A. Atkinson, Rev. Mod. Phys. 57 (1985) 437.

[8] H. Schmalzried, *Chemical Kinetics of Solids*, VCH (1995).

[9] M. Martin, Pure Appl. Chem. 75 (2003) 889.

[10] I. Barin, O. Knacke, *Thermochemical properties of inorganic substances*, Springer (1973).

[11] R. Merkle, Y. A. Mastrikov, E. A. Kotomin, M. M. Kuklja, J. Maier, J. Electrochem. Soc. 159 (2012) B219.

[12] H. J. M. Bouwmeester, H. Kruidhof, A. J. Burggraaf, Solid State Ionics 72 (1994) 185.

[13] H. S. Carslaw, J. C. Jaeger, *Conduction of Heat in Solids,* Clarendon Press, Oxford (1959); J. Crank, *Mathematics of Diffusion*, Clarendon Press, Oxford (1975).

[14] R. J. Chater, S. Carter, J. A. Kilner, B. C. H. Steele, Solid State Ionics 53 (1992) 859.

[15] M. Leonhardt, R. A. De Souza, J. Claus, J. Maier, J. Electrochem. Soc. 149 (2002) J19.

[16] F. S. Baumann, J. Fleig, H. U. Habermeier, J. Maier, Solid State Ionics 177 (2006) 1071.

[17] J. Claus, M. Leonhardt, J. Maier, J. Phys. Chem. Solids 61 (2000) 1199.

[18] J. A. Lane, S. J. Benson, D. Waller, J. A. Kilner, Solid State Ionics 121 (1999) 201.

[19] J. Jamnik, J. Maier, Phys. Chem. Chem. Phys. 3 (2001) 1668.

[20] L. Wang, R. Merkle, Y. A. Mastrikov, E. A. Kotomin, J. Maier, J. Mat. Res. 27 (2012) 2000.

[21] W. C. Chueh, A. H. McDaniel, M. E. Grass, Y. Hao, N. Jabeen, Z. Liu, S. M. Haile, K. F. McCarty, H. Bluhm, F. El Gabaly, Chem. Mater. 24 (2012) 1876.

[22] W. C. Jung, H. L. Tuller, Adv. Energ. Mater. 1 (2011) 1184.

[23] Y. L. Lee, J. Kleis, J. Rossmeisl, D. Morgan, Phys. Rev. B 80 (2009) 22401.

[24] Y. A. Mastrikov, R. Merkle, E. Heifets, E. A. Kotomin, J. Maier, J. Phys. Chem. C 114 (2010) 3017.

[25] R. Merkle, J. Maier, Phys. Chem. Chem. Phys. 4 (2002) 4140.

[26] R. A. De Souza, J. A. Kilner, Solid State Ionics 126 (1999) 153.

[27] A. T. Fromhold, J. Phys. Soc. Jap. 48 (1979) 2022.

[28] C. Wagner, Z. Phys. Chem. B21 (1933) 25.

[29] N. Cabrera, N. F. Mott, Rep. Prog. Phys. 12 (1949) 163.

[30] M. Holzinger, J. Maier, W. Sitte, Solid State Ionics 94 (1997) 217.

[31] K. D. Kreuer, Chem. Mater. 26 (2014) 361.

[32] K. D. Kreuer, Ann. Rev. Mater. Res. 33 (2003) 333.

[33] A. Hauch, S. D. Ebbesen, S. H. Jensen, M. Mogensen, J. Mater. Chem. 18 (2008) 2331.

[34] J. O'M. Bockris, A. K. N. Reddy, M. Gamboa-Aldeco, *Modern Electrochemistry*, Kluwer (1998)

[35] J. Fleig, Phys. Chem. Chem.Phys. 7 (2005) 2027.

[36] J. Fleig, R. Merkle, J. Maier, Phys. Chem. Chem. Phys. 9 (2007) 2713.

[37] C. Wagner, Z. Electrochem. 60 (1954) 4.

[38] M. H. Hebb, J. Chem. Phys. 20 (1952) 185.

[39] I. Yokota, J. Phys. Soc. Jap. 16 (1961) 2213.

[40] J. Maier, J. Phys. Chem Sol. 46 (1985) 197.

[41] I. Riess, Solid State Ionics 91 (1996) 221.

[42] H. Timm, J. Janek, Solid State Ionics 176 (2005) 1131.

[43] D. Monceau, M. Filal, M. Tebtoub, C. Petot, G. Petot-Evas, Solid State Ionics 73 (1994) 221.

[44] H. Schmalzried, W. Laqua, P. L. Lin, Z. Naturf. A34 (1979) 192.

[45] A. R. Allnatt, A. V. Chadwick, Trans. Farad. Soc. (1966) 1726.

[46] H. Timm, *PhD thesis*, University of Hannover, Germany (1999).

A 7 Electron Transport: Disorder and Correlations

Matthias Wuttig
I. Institute of Physics (IA) RWTH Aachen and
Forschungszentrum Jülich GmbH

Contents

1 Goal of this Lecture **2**

2 The Development of Transport Theory and the Boltzmann Equation **3**
2.1 Charge Transport Properties 3
2.2 Historical Background of Charge Transport Theory 5
2.3 The Boltzmann Equation 8
2.4 Success Stories of the Boltzmann Equation 10

3 Validity of the Boltzmann Equation **12**
3.1 Limits of the Boltzmann Equation 12
3.2 The Ioffe-Regel Criterion 14

4 Quantum Corrections to Conductivity **18**
4.1 Scattering mechanisms 18
4.2 Weak Localization 19
4.3 Effect of a Magnetic Field on Weak Localization 25
4.4 Summary 28

5 The Interplay of Disorder and Interactions **29**
5.1 Effect of Disorder on the Density of States 29
5.2 Effect of Electron-Electron Interactions on Transport 31
5.3 Determination of the Density of States by Tunneling Spectroscopy 32
5.4 Summary 34

6 Metal-Insulator Transition **36**
6.1 MIT pathways 36

References **44**

Lecture Notes of the 47$^{\text{th}}$ IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Goal of this Lecture

This lecture is devoted to electron transport. We are going to present a modern understanding of electron transport, which is based on the Boltzmann equation, but also contains a detailed discussion of the limits of applicability of transport theory based on the Boltzmann equation. In particular, you will learn when this equation can *not* be applied and when it has to be applied with great care, *i.e.* when additional correction terms have to be considered. In the end this should lead to a detailed understanding of the role that disorder and electron correlations play for charge transport. We will see how the interplay of disorder and correlations impacts charge transport and discuss recent research activities. Hence, this lecture series should also help you to link between textbook physics discussing scientific mysteries solved in the past and scientific questions that we are currently tackling. Hopefully, you will find these goals and this content both scientifically rewarding and interesting.

2 The Development of Transport Theory and the Boltzmann Equation

Outline

The present chapter aims to provide a quick overview of the basic concepts of electron transport theory usually encountered in a first course on solid-state physics. These concepts were developed between roughly 1895 and 1930. In particular, the following topics are discussed:

- ◇ Criteria to determine whether a material is a metal or an insulator
- ◇ Concept of metal-insulator transition (MIT)
- ◇ Historical development of the microscopic theory of electrons in solids
- ◇ Boltzmann equation to describe electron transport in a solid
- ◇ Success stories for the Boltzmann transport equation.

2.1 Charge Transport Properties

The crucial quantity describing charge transport is the conductivity σ, that relates the current density j to the electric field $\mathcal{E}$ across the material through the relation $j = \sigma\mathcal{E}$ and depends upon charge carrier concentration n and carrier mobility μ according to the relation $\sigma = e\mu n$ (for single channel transport), where e is the elementary charge.

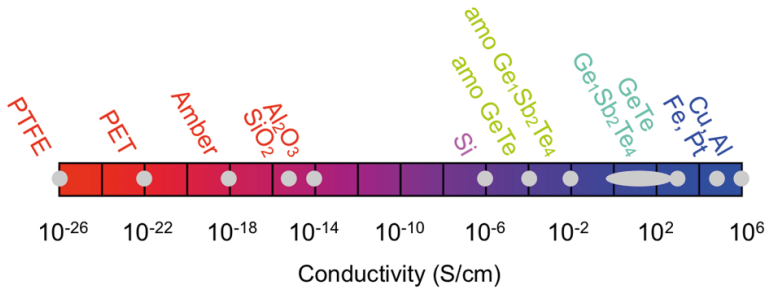
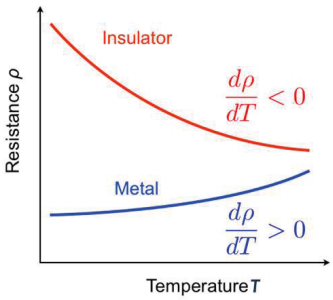


Figure 1: **Conductivity σ of various materials at room temperature $T = 300\,\mathrm{K}$:** The conductivity spreads over 32 orders of magnitude, whereas the atomic spacing only varies between 2 Å and 5 Å. (The data has been gathered by [1].)

In figure 1, the conductivity at room temperature ($T = 300\,\mathrm{K}$) of various materials is displayed. While the conductivity varies over an incredibly wide range of 32 orders of magnitude, a typical nearest neighbor distance between the atoms in a solid only ranges from 2 Å to 5 Å. The significant difference in conductivity enables to group materials in two different classes, *metals* and *insulators*, with two criteria to decide whether a material belongs to the former or the latter class.

1. Slope of the resistivity $\rho(T)$:

$$\frac{\partial \rho}{\partial T} < 0 \quad \Rightarrow \quad \text{insulator}$$
$$\frac{\partial \rho}{\partial T} > 0 \quad \Rightarrow \quad \text{metal}$$



N.B. Unfortunately and interestingly, this cannot be considered as an unambiguous criterion to distinguish between metals and insulators, since also metals can feature $\frac{\partial \rho}{\partial T} < 0$ *(see section 3.1 and 4.2).*

2. Low temperature limit of the conductivity:

$$\sigma\,(T \to 0) = 0\,K \quad \Rightarrow \quad \text{insulator}$$
$$\sigma\,(T \to 0) \neq 0\,K \quad \Rightarrow \quad \text{metal}$$

Bet: It is not possible to treat an elemental metal in a way that it will no longer show metallic behavior (i.e. a finite conductivity for $T \to 0\,K$*) in its solid state, no matter how many defects you induce. For the first counter example, a bottle of champagne is offered.*

The the bet on the conductivity of elemental metals should lead the reader to wonder about the possibility to turn some metals into insulators and vice versa by means of proper "treatments".

Vanadium dioxide (VO_2) is an example of a material undergoing a metal-insulator transition, as shown in figure 2: while the Hall mobility slightly changes at the transition temperature $T \approx 68\,°C$, the variation of the number of charge carriers per vanadium atom exeeds 4 orders of magnitude. The transition from a semiconducting transparent phase to a conducting non-transparent one and the possibility to reduce the transition temperature to $29\,°C$ by doping VO_2 with 1.9 at.% tungsten [3] enable the application in the architectural glazing industry. Indeed, architectural glass coated with a thin tungsten doped VO_2 layer can be used as an "intelligent" window glass which decreases the transmission of infrared light on warm days, but allows heat radiation to be transmitted into the house on cold days. Another application could be to coat the metal of a flat-iron with a material that undergoes an MIT at a temperature around $58\,°C$. The correlated change of optical properties could then be used as an indicator for the flat-iron being hot, so that one knows that it is too hot to touch it.

In VO_2, the MIT is based on a change of the crystal structure that induces a change in the electronic properties of the material; indeed, the material passes from a low temperature monoclinic phase with insulating properties to an high temperature rutile phase with metallic properties.

However, even more interesting is the origin of MIT without any change in the crystal structure. There are two different ways to have an electronically driven MIT:

1. **doping**, postulated by Sir Nevill Francis Mott;

2. **disorder**, postulated by Philip Warren Anderson.

Mott and Anderson - together with John van Fleck - were awarded the Nobel Prize in 1977 *"for their fundamental theoretical investigations of the electronic structure of magnetic and disordered systems"* [4]. Both have worked for decades on charge transport in solids, but their views differed considerably both with regard to possible origins of the MIT as well as the pathway to the insulating state. This will be discussed in section 6.

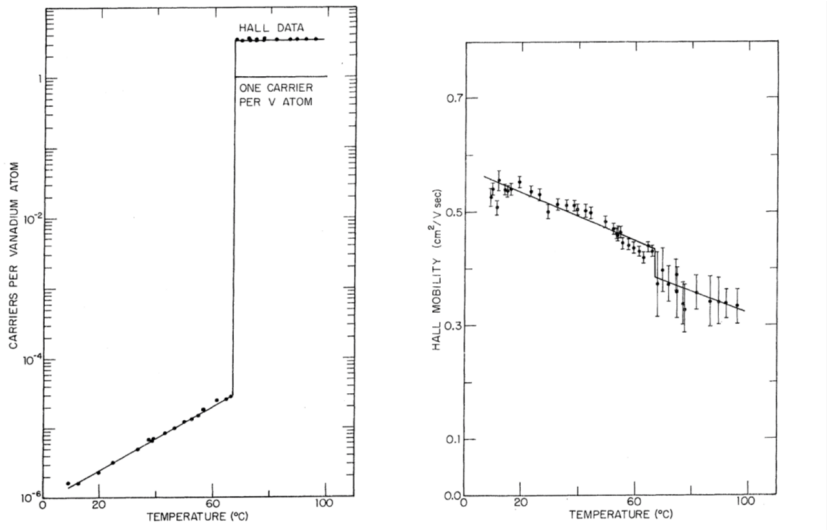


Figure 2: **Charge carriers and Hall mobility in VO_2 upon MIT:** VO_2 undergoes a metal-insulator transition at temperature $T \approx 68\,^{\circ}\mathrm{C}$. Upon this transition, the Hall mobility hardly changes, while the number of charge carriers per vanadium atom does change by more than 4 orders of magnitude. [2]

2.2 Historical Background of Charge Transport Theory

Drude's model of electron transport The first important development of a *microscopic* theory of charge transport came from Paul Drude in 1900, only three years after J. J. Thompson had discovered the electron (Nobel Prize in Physics 1906). Drude tried to build a theory in the spirit of the kinetic gas theory and assumed that electrons can be treated as classical particles with thermal energies that scatter at the ion cores in a solid. In an external field $\mathcal{E}$, the electrons are accelerated along the path between two collisions. With τ being the average time between two collisions, the equation of motion for an electron is given by

$$m\dot{v} + \frac{m}{\tau} v_{\mathrm{D}} \quad = \quad -e\mathcal{E}$$

where m is the electron mass, $v = v_{\mathrm{D}} + v_{\mathrm{therm}}$ is the velocity of an electron, v_{D} is the drift velocity, v_{therm} is the thermal velocity, e is the elementary charge.

It's worth to focus for a moment on the term τ. If we switch off the electric field, the electron velocity *relaxes* exponentially to the thermal velocity with a time constant τ; the *relaxation time approximation* is widely used in the study of transport phenomena, since it allows to combine good description and equation manageability.

In the stationary case ($\dot{v} = 0$), the drift velocity v_{D} becomes

$$v_{\mathrm{D}} \quad = \quad -\frac{e\tau}{m}\mathcal{E}.$$

By reminding that $v_D = -\mu\mathcal{E}$ for electrons, the electrical conductivity σ is thus given by

$$\begin{aligned} \sigma &= \frac{j}{\mathcal{E}} \\ &= \frac{-env_{\mathrm{D}}}{\mathcal{E}} \\ &= \frac{ne^2\tau}{m} \end{aligned} \tag{1}$$

where j is the current density, n is the electron density.
However, Drude made some severe mistakes when deriving Ohm's law:

1. treatment of electrons as classical particles, although they have to be treated as waves under specific conditions, as first shown by Sir G. P. Thompson (as his father J. J. Thompson before, Thompson Jr. was awarded with the Nobel Prize in Physics (in 1937));
2. electron energies of $k_{\mathrm{B}}T$;
3. scattering at ion cores in the crystalline solid (electrons are characterized by stationary states in a perfect periodic potential, thus no scattering events occur in absence of "disturbances" or boundaries, as we will see in section 2.2);
4. electron-electron interaction is neglected.

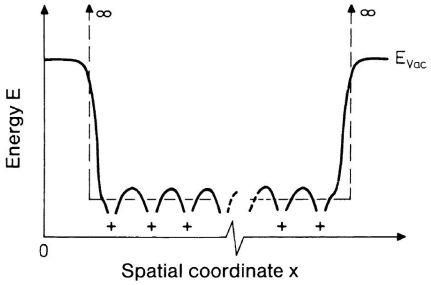


Figure 3: **Sommerfeld's approach to electron theory of metals: a)** Arnold Sommerfeld. **b)** Qualitative potential for an electron in a periodic lattice (continuous line) and approximation proposed by Sommerfeld in his model (dashed line).

Sommerfeld's theory of metals In 1933, Arnold Sommerfeld and Hans Bethe (Nobel Prize in Physics in 1967 *"for his contributions to the theory of nuclear reactions, especially his discoveries concerning the energy production in stars"* [5]) developed the model of the free electron gas in an infinite square-well potential to describe conduction electrons in metal (figure 3); electrons in the most external shell are hardly sensitive to the region close to the nuclei because of the effect of the core electrons, therefore the *effective potential* acting on them becomes a smoothly varying quantity that can be approximated by a constant.
In this model, the possible energy states are given by

$$E = E_0 + \frac{\hbar^2 k^2}{2m} = \frac{\hbar^2}{2m}\left(k_x^2 + k_y^2 + k_z^2\right),$$

where the metal crystal had been simplified to a cube of length L with infinite potential barrier at the surfaces; for simplicity, $E_0 = 0$ is usually assumed.
The components of the wave vector $k_x,\ k_y,\ k_z$ can be constrained by imposing that the wave function vanishes at the borders of the cube; it results

$$\begin{aligned} k_x &= \frac{\pi}{L} n_x, \\ k_y &= \frac{\pi}{L} n_y, \\ k_z &= \frac{\pi}{L} n_z, \end{aligned}$$

with $n_x, n_y, n_z = 1, 2, 3, ...$ With this limitation, only $\vec{k}$ values in the positive octant in $\vec{k}$-space are possible, which leads to the density of states (*cf.* [6], Sections 6.1, 6.2)

$$D(E) = \frac{(2m)^{3/2}}{2\pi^2\hbar^3} E^{1/2}. \tag{2}$$

With the number of occupied states

$$n = \int_0^\infty D(E) f(T, E) \mathrm{d}E$$

one obtains for the Fermi energy E_F^0 at $T = 0\,\mathrm{K}$

$$E_\mathrm{F}^0 = \frac{\hbar^2}{2m} \left(e\pi^2 n\right)^{2/3} \propto n^{2/3}.$$

The higher the density of electrons in a solid, the higher is the Fermi energy E_F^0.

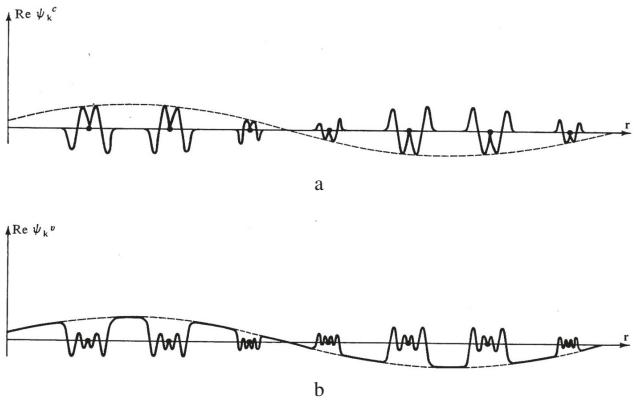


Figure 4: **Characteristic spatial dependence of Bloch wave functions:** (a) Core wave function and (b) valence wave function. The envelope is sinusoidal, *i.e.* a plane wave. [7]

Bloch's theory of electrons in solids Considering the periodic potential of the ion cores in a crystalline solid, the solutions of the Schrödinger equation becomes Bloch waves, named after Felix Bloch:

$$\psi_{\vec{k}}(\vec{r}) \;=\; u_{\vec{k}}(\vec{r}) \cdot \exp\left(i\vec{k}\cdot\vec{r}\right), \tag{3}$$

with the lattice-periodic modulation factor $u_{\vec{k}}(\vec{r}) = u_{\vec{k}}(\vec{r}+\vec{r}_n)$; the envelope of the Bloch waves is a plane wave (see figure 4). It's possible to prove that $\psi_{\vec{k}} = \psi_{\vec{k}+\vec{g_n}}$ and $E(\vec{k}) = E(\vec{k}+\vec{g_n})$, where g_n is the reciprocal lattice vector; therefore, by virtue of the Pauli's exclusion principle, it's common practice to narrow the analysis to the states within a "period" of the reciprocal lattice (ex. the first Brillouin zone). Drude had assumed scattering of the electrons at the periodic potential of positive ion cores (without the scattering, there could not be any electrical resistance). However, the Bloch waves that solve the stationary Schrödinger equation describe propagation without any scattering, since $\psi^*\psi$ is time-independent. For wave packets consisting of electron waves, which describe localized electrons, $\psi^*\psi$ is still time-independent. It follows that electrons cannot scatter with the ion cores.
Two are the possible ways how deviations from the undisturbed propagation can occur:

1. Electron scattering due to deviations from the full periodicity of the crystal lattice:
   (a) time-independent lattice defects such as dislocations and impurities;
   (b) time-dependent deviations from the periodicity, *i.e.* lattice vibrations.
2. Electron-electron scattering.

The Boltzmann equation takes such scattering events into account.

2.3 The Boltzmann Equation

The Boltzmann equation is a semi-classical approach to the problem of transport due to intra-band electronic processes, consequentially suitable for metals. In particular, the Boltzmann equation describes the influence of external fields and scattering events on the distribution function $f(\vec{r},\vec{k},t)$ of charge carriers.
Without any external fields and without any temperature gradient, the distribution function is given by the Fermi-Dirac distribution:

$$f_0\left[E\left(\vec{k}\right)\right] \;=\; \frac{1}{\exp\left[\frac{E\left(\vec{k}\right)-E_{\mathrm{F}}}{k_{\mathrm{B}}T}\right]+1} \tag{4}$$

In the presence of external fields and/or a temperature gradient, the non-equilibrium distribution $f\left(\vec{r},\vec{k},t\right)$ depends on space and time. By treating the effect of the applied fields on the distribution function classically and applying the Liouville's theorem, it's possible to write:

$$f(\vec{r}+\vec{v}dt,\vec{k}+\frac{\vec{F}}{\hbar},t) = f(\vec{r},\vec{k},t) + \left(\frac{\partial f}{\partial t}\right)_{col} dt \tag{5}$$

where $f = f(\vec{r},\vec{k},t)$ is the non-equilibrium distribution in the presence of external fields, $\vec{v} = \frac{1}{\hbar}\vec{\nabla}_{\vec{k}}E\left(\vec{k}\right)$ is the group velocity of an electron, and $\vec{F} = -q\left(\vec{\mathcal{E}} + \frac{1}{c}\vec{v}\times\vec{B}\right)$ is the force due to the external electric and magnetic field, respectively.

Taylor expansion of the term on the left side of the equation stopped at the first order leads to the Boltzmann transport equation:

$$\frac{\partial f}{\partial t} + \vec{v} \cdot \vec{\nabla}_{\vec{r}} f - \frac{e}{\hbar}\left(\vec{\mathcal{E}} + \frac{1}{c}\vec{v} \times \vec{B}\right) \cdot \vec{\nabla}_{\vec{k}} f \quad = \quad \left(\frac{\partial f}{\partial t}\right)_{\text{col}} \tag{6}$$

The semiclassical model describes the case displayed in figure 5: the wavelength of the applied field or temperature variation is larger than the spread of the wave package of the charge carrier [1], which again is larger than the spacing between two atoms.

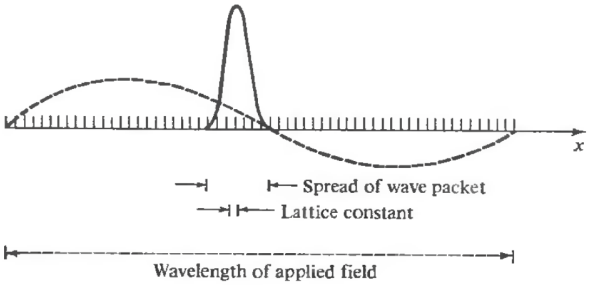


Figure 5: **Dimensions in the semiclassical model:** The wavelength of the applied field or temperature variation (dashed line) is larger than the spread of the wave package of the charge carrier (solid line), which again is larger than the atomic distance. [7]

The terms on the left, that describe the influence of external fields on the distribution of charge carriers, are called "drift terms"; the term on the right is called "collision term" and takes the effect of scattering events into account.
The complexity of equation 6 is caused by the collision term. Often, one simplifies this term by using the "relaxation time approximation", already mentioned in the context of the Drude model. It is assumed that the distribution function $f\left(\vec{r}, \vec{k}, t\right)$ relaxes to its equilibrium $f_0\left(\vec{r}, \vec{k}, t\right)$ within the time $\tau\left(\vec{r}, \vec{k}\right)$:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{col}} \quad = \quad -\frac{f\left(\vec{r}, \vec{k}, t\right) - f_0\left(\vec{r}, \vec{k}, t\right)}{\tau\left(\vec{r}, \vec{k}\right)}. \tag{7}$$

The conductivity of a metal is given by

$$\sigma \quad \backsimeq \quad \frac{e^2 \tau\left(E_{\text{F}}\right)}{m^*} n \tag{8}$$

where $m^* = \hbar^2 \left(\frac{\mathrm{d}^2 E}{\mathrm{d}k^2}\right)^{-1}$ is the effective mass of a quasi free electron under the parabolic band approximation.

[1] The electrons in a solid are described by Bloch waves (*cf.* [6], Section 7.1).

Equation 8 is very similar to Drude's formula (*cf.* equation 1), but here it has been taken into account that only electrons near the Fermi edge E_F can contribute to an electric current and instead of the electron mass m the effective mass m^* of the electron in the solid is employed.
For small electron scattering rates, one can assume that the rates for different scattering mechanisms can be added:

$$\frac{1}{\tau} = \frac{1}{\tau_{\mathrm{e-def}}} + \frac{1}{\tau_{\mathrm{e-ph}}} + \frac{1}{\tau_{\mathrm{e-e}}} \tag{9}$$

where $\tau_{\mathrm{e-def}}$ is the time constant for elastic scattering of electrons with defects, $\tau_{\mathrm{e-ph}}$ is the time constant for inelastic scattering of electrons with phonons, and $\tau_{\mathrm{e-e}}$ is the time constant for scattering of electrons with each other.
In most cases, electron-electron scattering can be neglected. The cross section of scattering is reduced not only due to the screening effect, but also due to the small number of electrons being able to contribute to this scattering mechanism: only electrons in a range $2k_\mathrm{B}T$ around the Fermi edge E_F can scatter into unoccupied states in $\vec{k}$ space. Since this argument holds for both electrons involved, the cross section is reduced by $\left(\frac{T_\mathrm{F}}{T}\right)^2$. With typical values of $E_\mathrm{F} \sim 10^5\,\mathrm{K}$, even at room temperature we only expect a cross section

$$\Sigma \propto \left(\frac{k_\mathrm{B}T}{E_\mathrm{F}}\right)^2 \Sigma_0 \propto \left(\frac{T}{T_\mathrm{F}}\right)^2 \Sigma_0 \sim \left(\frac{3\cdot 10^2}{10^5}\right)^2 \Sigma_0 = 10^{-5}\,\Sigma_0.$$

Using equation 9 one obtains Matthiessen's rule, which states that the different scattering mechanisms contribute additively to the resistance $\rho = 1/\sigma \propto 1/\tau$:

$$\begin{aligned} \rho(T) &= \rho_\mathrm{def} + \rho_\mathrm{ph}(T) + \rho_{\mathrm{e-e^-}}(T) \\ &\approx \rho_\mathrm{def} + \rho_\mathrm{ph}(T). \end{aligned} \tag{10}$$

While ρ_def does not depend on the temperature, $\rho_\mathrm{ph} = \rho_\mathrm{ph}(T)$ does: $\rho_\mathrm{ph}(T > \Theta) \propto T$ (where Θ is the Debye temperature), while Grüneisen found $\rho_\mathrm{ph}(T \ll \Theta) \propto T^5$ (*cf.* [6], Section 9.5). In particular, at $T = 0\,\mathrm{K}$ there are no lattice vibrations, so that $\rho(T \to 0\,\mathrm{K}) \to \rho_\mathrm{def}$.

2.4 Success Stories of the Boltzmann Equation

The temperature dependence of scattering events is clearly visible in measurements of resistance R versus temperature T for three samples of sodium with different defect concentrations (see figure 6). Figure 7 displays the resistance measured as a function of temperature between $100\,\mathrm{K}$ and $300\,\mathrm{K}$ for copper and copper-nickel alloys of different compositions. For all alloys the linear temperature dependence of the resistance due to phonon scattering is clearly visible. With increasing nickel concentration, *i.e.* with an increase of impurities, the resistance curve is shifted vertically due to the additive contribution of ρ_def.
Plotting the reduced resistivity R/R_Θ (with $R_\Theta = R(T = \Theta)$) versus the reduced temperature T/Θ validates Matthiessen's rule (equation 10) for a variety of metals (see figure 8).
Up to now, the Boltzmann equation seems to have a great success. However, in the following section 3.1 we will see that besides those wonderful successes there are also significant failures. This will lead us to a "traffic light" in section 3.2 illustrating in which case the Boltzmann equation can be used without any restrictions, in which case additional quantum mechanical terms have been considered, and in which case the Boltzmann equation is invalid.

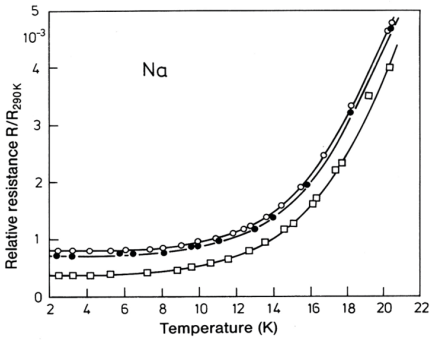


Figure 6: **Matthiessen's rule, using the example of sodium:** The electrical resistivity R – normalized to the resistivity $R_{290\,\mathrm{K}}$ at $T = 290\,\mathrm{K}$ – is plotted against the temperature T for three samples of different defect concentrations [8]. For $T < 8\,\mathrm{K}$ the constant contribution R_{def} from scattering at defects is visible. At higher temperatures the resistivity increases due to phonon scattering as described by Grüneisen. For $T > 18\,\mathrm{K}$, the resistivity increases linearly with temperature. Primary source: [8], secondary source: [6].

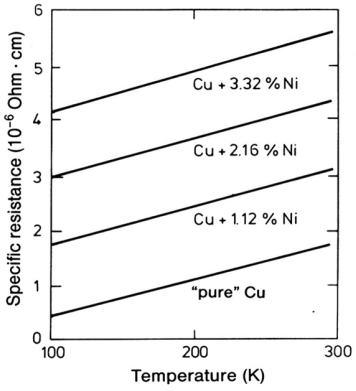


Figure 7: **Matthiessen's rule, using the example of different copper alloys:** Temperature dependence of the resistance ρ for copper and copper-nickel alloys of different compositions [9]. The linear contribution from phonon scattering is clearly visible between 100 K and 300 K. The nickel atoms serve as scattering centers. Due to the additive contribution of ρ_{def} the $\rho(T)$ curve is vertically shifted with increasing nickel concentration. Primary source: [9], secondary source: [6].

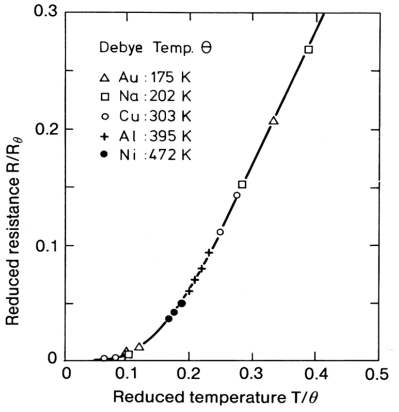


Figure 8: **Matthiessen's rule, using the example of different metals:** Plotting the reduced resistivity R/R_Θ versus the reduced temperature T/Θ validates Matthiessen's rule equation 10 for a variety of metals. [6]

3 Validity of the Boltzmann Equation

Outline

The resistivity phenomena introduced in section 2.4 could be nicely explained by the Boltzmann equation 6. Unfortunately, there are also phenomena that need further quantum mechanical terms or even fail to be explained by the Boltzmann equation.
In the following, we will see:

- ◇ examples for the deviation of the resistance behavior from the expected one based on the Boltzmann equation;
- ◇ a criterion for the applicability of the Boltzmann equation, called Ioffe-Regel criterion.

3.1 Limits of the Boltzmann Equation

J. H. Mooij from Philips Eindhoven plotted the experimental Temperature Coefficients of Resistance (TCR) $\bar{\alpha} = \frac{1}{R}\frac{\mathrm{d}R}{\mathrm{d}T}$ versus the resistivity ρ for many high-resistance alloys (see figure 9, [10]). Thereby, he could show that the slope of the resistance versus temperature decreases with increasing resistivity, establishing the so-called *Mooij rule*: alloys with resistivity below $(100 - 150)\,\mu\Omega$·cm feature a positive TCR, while alloys with resistivity above $(100 - 150)\,\mu\Omega$·cm show a negative TCR [11].
A decrease of the resistivity with temperature would correspond to the scattering time τ to increase with increasing temperature. No such scattering mechanism is known, therefore the Boltzmann equation fails to describe this phenomenon.

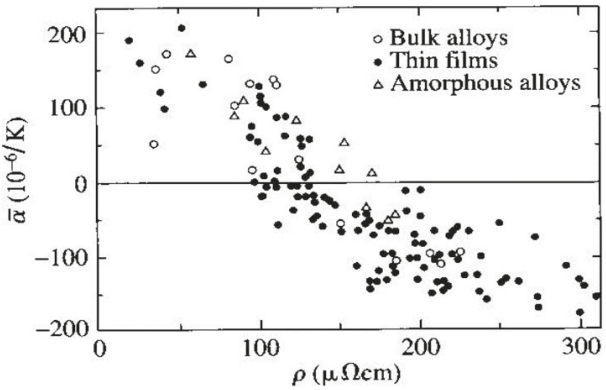


Figure 9: **Mooij rule:** Experimental temperature coefficients of resistance, $\bar{\alpha} = \frac{1}{R}\frac{\mathrm{d}R}{\mathrm{d}T}$, are plotted versus resistivity ρ for high-resistance alloys. The slope of $\bar{\alpha}$ decreases with increasing resistivity. Since no such scattering mechanism with negative $\mathrm{d}\tau/\mathrm{d}T$ is known, the Boltzmann equation fails to explain this phenomenon. Primary source: [10], secondary source: [11].

Another example for failure of the Boltzmann equation is displayed in figure 10 (primary source: [12], secondary source: [11]): Nb_3Sb and Nb_3Sn both are superconductors with transition temperature $T_C\,(Nb_3Sb) = 0.2\,K$ and $T_C\,(Nb_3Sn) = 18\,K$, respectively. The BCS theory

explains the formation of bound cooper pairs as a consequence of an attractive electron-electron interaction mediated by phonons. The transition temperature T_C is correlated to the phonon frequency (see for example [6], Section 10.6). In Nb_3Sn the contribution from electron-phonon scattering to the resistivity ρ is thus expected to be higher than in Nb_3Sb. Indeed, this is true for small temperatures, but above $500\,\mathrm{K}$ the resistivity $\rho(T \gtrsim 500\,\mathrm{K})$ is the same for both compounds. Furthermore, for both materials $\rho(T \gtrsim 500\,\mathrm{K})$ tends to saturate towards about $150\,\mu\Omega\cdot\mathrm{cm}$. This is not consistent with the linear contribution of electron-phonon scattering to the resistivity.

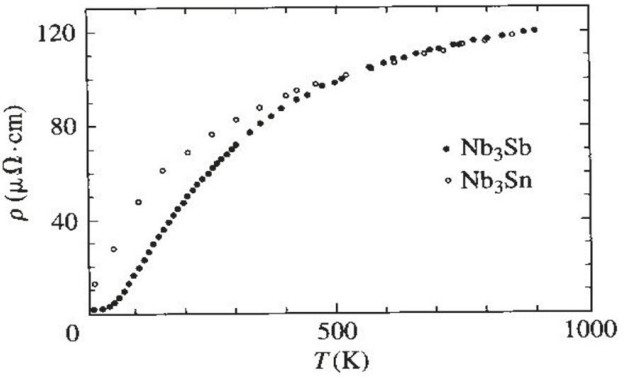


Figure 10: **Normal-state electrical resistivities for single-crystals of the superconductors Nb_3Sb and Nb_3Sn:** Both compounds feature a saturation of the resistivity ρ above temperatures $T \gtrsim 500\,\mathrm{K}$. Although their transition temperatures T_C significantly differ from each other ($T_C\,(Nb_3Sb) = 0.2\,\mathrm{K}$, $T_C\,(Nb_3Sn) = 18\,\mathrm{K}$), the saturation level is the same. This is not consistent with the linear contribution to the resistivity expected from electron-phonon scattering. Primary source: [12], secondary source: [11].

The phenomenon of saturation of $\rho(T)$ has also been experimentally shown for TiAl alloys (see figure 11). Already pure Ti shows this characteristic. As expected, with increasing Al concentration, $\rho(T \to 0\,\mathrm{K}) = \rho_{\mathrm{def}}$ increases.[2] However, the increasing Al concentration leads to a weaker resistivity increase with temperature [10, 11], so that the saturation levels for different TiAl alloys are in the same regime. An Al concentration of 33 % even leads to a negative slope $\frac{\partial\rho}{\partial T} < 0$. This is not consistent with the identification of a material as a metal or insulator via $\frac{\mathrm{d}\rho}{\mathrm{d}T} < 0$ and $\frac{\mathrm{d}\rho}{\mathrm{d}T} > 0$, respectively, as proposed on page 4, since the material still is clearly metallic due to the resistivity $\rho(T)$ being finite at $T = 0\,\mathrm{K}$.

Now we are in deep trouble: in section 2.4 we have seen wonderful successes of the Boltzmann equation, while in this section many materials have been shown for which the Boltzmann equation fails to explain the observed temperature dependence of the resistivity. We thus have to find a criterion to establish whether the Boltzmann equation can be used without any restrictions,

[2] We have seen before that with increasing amount of defects the resistivity is shifted linearly to higher values (*cf.* figure 7). This observation follows Matthiessen's rule (see equation 10) and the fact that the contribution of defects to the resistivity $\rho_{\mathrm{def}}(T) = \rho_{\mathrm{def}}$ does not depend on the temperature T.

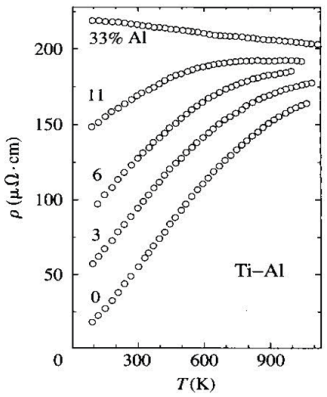


Figure 11: **Resistivities of different TiAl alloys:** With increasing Al concentration, the increase of the resistivity $\rho(T)$ with temperature becomes weaker. $Ti_{67}Al_{33}$ even shows a negative slope of resistivity $\frac{\partial\rho}{\partial T} < 0$. Primary source: [10], secondary source: [11].

if additional quantum mechanical terms have been considered or if the Boltzmann equation is invalid. This will be subject of the following section.

3.2 The Ioffe-Regel Criterion

In the previous chapter we have discussed charge transport in metals. The equation to use is the Boltzmann equation 6, which describes how the distribution function $f(\vec{v}, \vec{k}, t)$ is modified by an external electric field as well as by electron collisions. Such collisions include electron-electron collisions, which in most cases can be ignored in solids even though the electron density is very high. The two remaining relevant collision mechanisms are electron-defect and electron-phonon scattering. For small electron scattering rates we can add these three scattering channels (*cf.* Matthiessen's rule equation 10) and discuss the temperature dependence of the resistivity of metals (*cf.* page 2.3).
Indeed we have seen three examples of such $R(T)$ data in section 2.4 which can be nicely explained by the Boltzmann equation. We have even discussed an interesting challenge for the application of thin Ag films. Clearly understanding and applying the Boltzmann equation has proven useful. However, in section 3.1 we have also seen data where it was impossible to explain the resistivity using the Boltzmann equation. This implies that we get flawed (wrong) predictions if we apply this equation under conditions where this formula is not valid.
At this point we should think about a condition that tells us if the Boltzmann equation can be applied. The concept of the Boltzmann equation assumes electrons propagating as Bloch waves (3) and scattering with defects, phonons and other electrons (*cf.* section 2.3). The prerequisite for using this equation is the mean free path l of the electron being much larger than the Fermi wavelength $\lambda_F = 2\pi/k_F$, where $k_F = (3\pi^2 n)^{\frac{1}{3}}$ is the Fermi wavevector, defined by assuming spherical Fermi surface[3]. Neglecting numerical factors of the order of unity leads to

$$k_F \cdot l \gg 1 \tag{11}$$

where $k_F = 2\pi/\lambda_F$ is the Fermi wave vector (with λ_F being the distance between two adjacent maxima of the Bloch wave); l is the mean free path between two collisions.

[3]The Fermi surface is defined as the set of points in the reciprocal space such that $E(\vec{k}) = E_F$

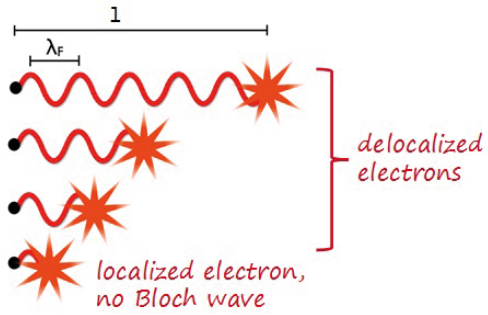


Figure 12: **Illustration of the Ioffe-Regel criterion:** The condition for using the Boltzmann equation is $k_F \cdot l = 2\pi l/\lambda_F \gg 1$.

Indeed, when $\lambda_F \sim l$ we can't properly talk about propagating waves anymore since the path travelled between two collisions becomes shorter than the wavelength (see 12).
The condition can be demonstrated according to quantum mechanical considerations as well: from the uncertainty principle we know that $\Delta k \Delta x \sim 1$; since $k_F \gg \Delta k$, it follow $\Delta x \gtrsim 1/k_F$, which combined to the fact that the distance l between two scattering centers should exceed the minimum uncertainty, implies: $k_F \times \gtrsim 1$. This relation is called *Ioffe-Regel criterion*.
Three situations can now be distinguished according to table 1. Chapter 4 will deal with the situation $k_F \cdot l \gtrsim 1$ of weak localization of electrons, where the Boltzmann equation works, but fails to describe certain features, so that additional quantum mechanical terms will be necessary. For $k_F \cdot l < 1$, the Boltzmann equation becomes invalid. Electrons do not behave like extended waves (Bloch states) anymore. They become localized This regime is calledstrong localization.

	$k_F \cdot l < 1$	strong localization	Boltzmann equation invalid
	$k_F \cdot l \gtrsim 1$	weak localization	Boltzmann equation valid at high T, quantum corrections at low T
	$k_F \cdot l \gg 1$	ordinary metal	Boltzmann equation valid

Table 1: **Limits of the validity of the Boltzmann equation**

Sir Nevill Francis Mott (Nobel Prize in 1977, see page 4) was convinced that something special should happen when $k_F \cdot l$ approaches 1; in particular, with increasing disorder the conductivity decreases till it reaches a minimum value $\sigma_{\min}$ after which it suddenly drops to 0. This phenomenon is called *minimum metallic conductivity* (MMC). On the other hand, Philip Warren Anderson (Nobel Prize in 1977, see page 4) was convinced of a continuous, second-order metal-insulator transition (MIT). The two situations are displayed in figure 13. But who was right? The phenomenon of the MIT will be the topic of section 6.
A criterion using $k_F \cdot l$ is not very obvious: researchers normally do not deal with the value $k_F \cdot l$. Instead, it is rather straightforward to use the transport properties as a function of temperature

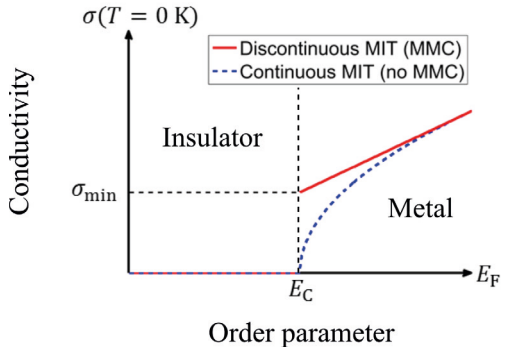


Figure 13: **The minimum metallic conductivity:** The conductivity decreases with increasing order and – according to Sir Nevill Francis Mott – reaches a minimum value σ_{min} at a critical value of disorder, after which the conductivity suddenly drops to 0 (red line). The minimum metallic conductivity (MMC) corresponds to the transition from metal to insulator. Other scientists like Philip Warren Anderson were convinced that the metal-insulator transition (MIT) takes place continuously (dashed blue curve). Primary source: [13], secondary source: [14].

(*cf.* page 4) to distinguish between metallic and non-metallic behavior. An MIT indicated by the change of the slope $\mathrm{d}\rho/\mathrm{d}T$ is displayed in figure 14: four different phase change alloys have been annealed to different temperatures and upon cooling their resistivity versus temperature data points $\rho(T)$ were measured [15]. At a a critical resistivity $\rho = (2-3)\,\mathrm{m\Omega{\cdot}cm}$ the slope of the $\rho(T)$ curves changes from negative, *i.e.* non-metallic behavior, to positive *i.e.* metallic behavior, while the charge carrier density n remains constant. (This critical resistivity is the same for all the for phase change alloys displayed in figure 14.) Interesting enough, this TCR sign change happens when $k_{\mathrm{F}} \cdot l$ equals 1.

For classification of a material regarding the value for $k_{\mathrm{F}} \cdot l$, it is sufficient to determine the resistivity ρ and the charge carrier density n from Hall measurements:

$$
\begin{aligned}
n &= M\frac{k_{\mathrm{F}}^3}{3\pi^2} && \text{(see [6], Section 6.2)} \\
\Rightarrow k_{\mathrm{F}} &= \left[\frac{3\pi^2 n}{M}\right]^{\frac{1}{3}} && (12) \\
\rho = \sigma^{-1} &= \left(\frac{ne^2\tau}{m^*}\right)^{-1} && \text{with } \tau = l v_{\mathrm{F}}^{-1} = l\left(\frac{\hbar k_{\mathrm{F}}}{m^*}\right)^{-1} \\
&= \frac{3\pi^2\hbar}{e^2}\frac{1}{k_{\mathrm{F}}^2 l M} \\
&= \frac{3\pi^2\hbar}{e^2}\frac{1}{lM}\left[\frac{3\pi^2 n}{M}\right]^{-\frac{2}{3}} \\
\Rightarrow l &= \left(\frac{3\pi^2}{M}\right)^{1/3}\frac{\hbar}{e^2\rho}\cdot n^{-\frac{2}{3}} && (13)
\end{aligned}
$$

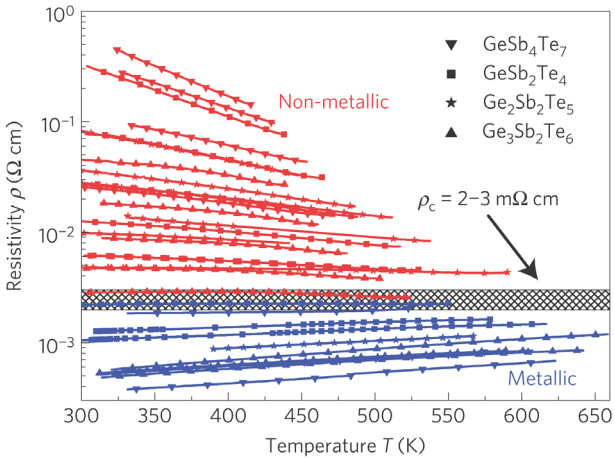


Figure 14: **Metal-insulator transition of four phase change alloys:** The resistivity versus temperature data points $\rho(T)$ of four different phase change alloys after having been annealed to different temperatures are displayed. All these alloys show a universal TCR sign change at a critical resistivity $\rho = (2-3)\,\mathrm{m\Omega \cdot cm}$: At this resistivity the slope of the $\rho(T)$ curves changes from negative ($\mathrm{d}\rho/\mathrm{d}T < 0 \rightarrow$ insulator) to positive ($\mathrm{d}\rho/\mathrm{d}T > 0 \rightarrow$ metallic behavior). [15]

$$\Rightarrow k_\mathrm{F} \cdot l = \left(\frac{3\pi^2}{M}\right)^{2/3} \frac{\hbar}{e^2\rho} \cdot n^{-\frac{1}{3}} \tag{14}$$

where v_F is the Fermi velocity; M is the the multiplicity of the valence band maximum.

Summarizing, in this chapter we got to know the Ioffe-Regel criterion, which states that when the mean free path l of the electron in a solid becomes as small as the Fermi wavelength $\lambda_\mathrm{F} = 2\pi/k_\mathrm{F}$, the propagation cannot be described by Bloch waves (3) any longer. The Boltzmann equation thus is only valid without any restrictions if $k_\mathrm{F} \cdot l \gg 1$. When the mean free path l becomes approximately of the order of the Fermi wavelength λ_F, *i.e.* $k_\mathrm{F} \cdot l \gtrsim 1$, additional correction terms to the conductivity have to be considered. This regime, which is called 'dirty metals' will be discussed in the following chapter. If the mean free path decreases further, so that $k_\mathrm{F} \cdot l < 1$, the Boltzmann equation becomes invalid.

4 Quantum Corrections to Conductivity

Outline

In the previous chapters, the description of electron transport in solids in terms of Bloch waves and the Boltzmann equation has been introduced. Furthermore, the Ioffe-Regel criterion, which enables to determine the applicability of the BTE has been presented.
While $k_F \cdot l \gg 1$ holds for a good metal, we now want to deal with metals where $k_F \cdot l \gtrsim 1$, i.e. metals where the scattering events happen more frequently and the Fermi wavelength and electron mean free path are comparable. As already mentioned, quantum corrections to the Boltzmann expression for the conductivity have to be incorporated into the model.
In particular, we will:

- ◇ briefly review the main scattering mechanisms in metals and define the diffusive transport regime;
- ◇ see how an electron can interfere with itself at low temperature, leading to weak localization;
- ◇ see how to destroy the weak localization regime by the application of a magnetic field.

4.1 Scattering mechanisms

Before we are going to focus on weak localization in the following section, different scattering mechanisms have to be distinguished:

- ◇ Elastic scattering:
  The electron scatters at stationary deviations from the perfectly periodic lattice, such as defects; the energy is conserved and the phase of the electron's Bloch wave encounters a constant change. The elastic mean free path l_e is given by $l_\mathrm{e} = v_\mathrm{F} \cdot \tau_\mathrm{e}$.
- ◇ Inelastic scattering:
  The electron scatters at non-stationary deviations from the periodicity, such as phonons and electrons; the energy is not conserved and the phase of the wave is changed stochastically. The inelastic mean free path l_in is given by $l_\mathrm{in} = \sqrt{D\tau_\mathrm{in}}$, where $D = \frac{l_\mathrm{e} v_\mathrm{F}}{d}$[4] is the diffusion coefficient and d is the dimensionality.
- ◇ Spin-flip scattering:
  The spin of the electron is reversed by a scattering event. The energy might not be conserved and the phase information of the original wave might be lost.

The *phase-breaking length*, after which the phase is changed stochastically, is given by $l_\varphi = \sqrt{D\tau_\varphi}$, and clearly depends upon inelastic and spin flip scattering events. Without any spin-flip scattering events, $l_\varphi = l_\mathrm{in}$. (*Cf.* [6], Section 9.9.)
In a metal, the regime of *diffusive transport* occurs $l_{in} \gg l_e$, i.e. when elastic scattering events are predominant; the condition is satisfied in "dirty" metals, i.e. with many defects (small l_e), at low temperature (large l_{in}).

[4] For mainly elastic scattering $D \approx \frac{l_e^2}{\tau_\mathrm{e} \cdot d}$.

4.2 Weak Localization

Analysis of the phenomenon Let's consider a dirty metal ($k_F \cdot l \gtrsim 1$) at low temperature, i.e. $l_e \ll l_{in}, l_\phi$, without any spin-flip scattering event ($l_\phi = l_{\text{in}}$). The transport is diffusive and we can write:

$$k_F \cdot l \approx k_F \cdot l_e \gtrsim 1, \tag{15}$$

where

$$\frac{1}{l} = \sum_i \frac{1}{l_i} \approx \frac{1}{l_e}. \tag{16}$$

In the diffusive transport regime, the electron realizes a random walk with the elastic mean free path $l_{\text{e}} = v_{\text{F}} \cdot \tau_{\text{e}}$ and the probability $p(\vec{r}, t)$ to find a diffusing particle after time t at a point $\vec{r}$ is given by the Gaussian distribution:

$$p(\vec{r}, t) \;=\; (4\pi D t)^{-\frac{d}{2}} \exp\left(-\frac{r^2}{4Dt}\right) \quad \text{, with } r^2 \;=\; \sum_i^d x_i^2 \text{ and } \int p(\underline{r}, t) d\underline{r} \;=\; 1 \tag{17}$$

where again

d is the dimensionality of the space in which diffusion takes place;

D is the diffusion coefficient $D = \frac{l_{\text{e}} v_{\text{F}}}{d}$.

The distribution Δr of the particle's location gradually becomes wider with time t:

$$\Delta r \;\simeq\; \sqrt{Dt} \;\approx\; \sqrt{\frac{l_{\text{e}}^2}{\tau_{\text{e}} \cdot d} \cdot t} \;\simeq\; l_{\text{e}} \sqrt{\frac{t}{\tau_{\text{e}}}} \;\simeq\; l_{\text{e}} \sqrt{N} \tag{18}$$

where $N = \frac{t}{\tau_{\text{e}}}$ is the number of steps in the diffusion process, *i.e.* the number of *elastic* scattering events within time t.

These equations describe a *classical* particle. However, care must be taken, since the wave properties of an electron have a significant impact on the probability density function $p(\vec{r}, t)$ at the origin $\vec{r} = 0$, leading to an increase of its amplitude at the origin.

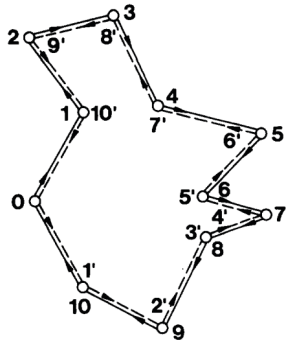


Figure 15: **Two possible paths of a diffusing electron that returns to its origin:** For classical diffusion the probability of one direction equals that one of the opposite direction. However, the electron has wave-like character. The probability of backscattering is increased for the quantum mechanical calculation. [16]

This phenomenon can be easily understood by considering the two dimensional case in figure 15, where two possible paths with same route but opposite direction for the return of the

electron to the origin are considered. As long as only elastic scattering events occur, the phases of the two partial waves propagating on the same path in opposite directions are coherent at the origin and the amplitudes are equal.
Upon classical considerations, the probability for an electron to propagate on one path equals the probability for the opposite direction and the probability for returning to the origin is obtained by adding the intensities of the waves:

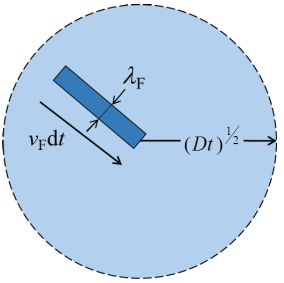
$$p(\vec{r}=0,t)_{\text{classical}} = |A_1|^2 + |A_2|^2 \overset{A_1=A_2}{=} 2A_1^2 = \frac{1}{4\pi Dt}. \tag{19}$$

However, due to the wave-like character of electrons, the two partial waves of electrons that propagate in opposite directions back to the origin constructively interfere with one another, resulting is a backscattering probability that is twice as large as the one predicted upon classical considerations:

$$\begin{aligned} p(\vec{r}=0,t)_{\text{qm}} &= |A_1 + A_2|^2 = |A_1^2| + |A_2^2| + 2|A_1 A_2| \overset{A_1=A_2}{=} 4A_1^2 \\ &= 2 \cdot p(\vec{r}=0,t)_{\text{classical}} \\ &= \frac{1}{2\pi Dt}. \end{aligned} \tag{20}$$

This significant constructive interference only takes place in the vicinity of the origin, since at any other point the two waves are generally not coherent. Since the constructive interference of partial electron waves at the origin leads to an increased probability to find the electron at $\vec{r} = 0$, this quantum diffusion phenomenon is called *weak localization* (a localized electron would remain close to the origin).
In case of spin-orbit scattering events, the partial waves are no longer coherent at the origin because of phase information loss and constructive interference no longer takes place; this results in reduction of the backscattering probability, leading to *weak antilocalization*.
The impact of weak localization and antilocalization on the probability distribution function is depicted in (figure 16).

Effect on transport properties: quantitative analysis In the following, the quantitative impact of weak localization on the electrical conductivity will be determined. For this purpose, information about the number of electrons that are able to revisit the origin within time $t < \tau_\varphi$, *i.e.* before the phase information is lost, is necessary.
Let's consider the three-dimensional case ($d = 3$). The volume where the electron can be found at time t is of the order of $(Dt)^{3/2}$.[5] The volume from which an electron can reach the origin within the time dt is given by $v_\text{F}\lambda_\text{F}^2 \text{d}t$. The relative number of electrons that are able to revisit the origin within time dt is then given by the ratio of these volumes, $v_\text{F}\lambda_\text{F}^2 \text{d}t / (Dt)^{3/2}$.



The minimum time for electrons to revisit the origin is given by the elastic scattering time τ_e. Those electrons can interfere constructively with themselves if they have not changed their phase yet, i.e. if they return to the origin before the phase breaking time ($\tau < \tau_\phi$).

[5] This can be estimated by calculating the volume $\frac{4}{3}\pi l^3 = \frac{4}{3}\pi \left(\sqrt{Dt}\right)^3 \approx (Dt)^{3/2}$.

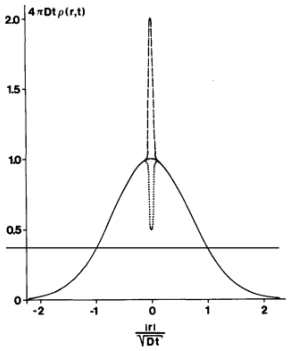


Figure 16: **Classical and quantum-mechanical probability distribution of a diffusing electron:** The electron starts at point $\vec{r} = 0$ and time $t = 0$. The quantum-mechanical description of the electron leads to constructive interference at the origin (dashed line), where the phases of the two partial waves are coherent without any inelastic, phase-breaking scattering events. This probability for backscattering is twice as large as in the classical description (full curve). Spin-orbit scattering events, on the other hand, reduce the probability for backscattering by a factor of two (dotted peak). [16]

The quantum correction to the conductivity due to weak localization in the three-dimensional case ($d = 3$) then is given by

$$\frac{\partial\sigma_{d=3}}{\sigma} \simeq -\int_{\tau_e}^{\tau_\varphi} \frac{v_F \lambda_F^2}{(Dt)^{3/2}}\,\mathrm{d}t \tag{21}$$

$$= -\frac{v_F\lambda_F^2}{D^{3/2}}\left(\frac{1}{\tau_e^{1/2}} - \frac{1}{\tau_\varphi^{1/2}}\right)$$

$$= -\frac{4\pi^2 v_F}{k_F^2 D}\left(\frac{1}{\sqrt{D\tau_e}} - \frac{1}{\sqrt{D\tau_\varphi}}\right)$$

$$= -\frac{4\pi^2 v_F d}{k_F^2 l_e v_F}\left(\sqrt{\frac{d}{l_e v_F \tau_e}} - \frac{1}{l_\varphi}\right)$$

$$= -\frac{12\pi^2}{k_F^2}\frac{1}{l_e}\left(\frac{\sqrt{d}}{l_e} - \frac{1}{l_\varphi}\right)$$

$$\simeq -\frac{1}{k_F^2 l_e}\left(\frac{1}{l_e} - \frac{1}{l_\varphi}\right). \tag{22}$$

The negative sign comes from the *increased* probability of backscattering, which leads to an increase of the resistivity and thus to a decrease of the conductivity. The phase-breaking length l_φ is much larger than the elastic mean free path l_e:

$$l_\varphi = \sqrt{D\tau_\varphi} \qquad \text{with } D = \frac{l_e^2}{\tau_e d}$$

$$
\begin{aligned}
&= l_{\mathrm{e}}\sqrt{\frac{\tau_\varphi}{\tau_{\mathrm{e}}}}\cdot\frac{1}{\sqrt{d}}\\
&\simeq l_{\mathrm{e}}\sqrt{\frac{\tau_\varphi}{\tau_{\mathrm{e}}}}\\
&\gg l_{\mathrm{e}}\,.
\end{aligned}
$$

Please note:

- ◇ The quantum correction to the conductivity in three dimensions is linear in $(k_{\mathrm{F}}l_{\mathrm{e}})^{-1}$ (1^{st} order perturbation theory). This correction works well if $k_F l_e \gg 1$, but becomes problematic if $k_F l_e$ approaches 1.
- ◇ The quantum correction is relatively small.
- ◇ The correction depends on temperature since – other than the elastic mean free path l_{e} – the phase-breaking length $l_\varphi = l_\varphi(T)$ depends on the temperature. Thus, the temperature dependence of the quantum correction to conductivity comes from inelastic, phase-breaking scattering events. As $T \to 0$, τ_φ tends to infinity, so that quantum corrections become more significant.

Although diffusion takes place as a result of elastic scattering with mean free path l_{e}, for the decision about dimensionality the comparison between the sample size a and the phase-braking length l_φ is crucial. While at high temperatures T the phase-breaking length l_φ is short so that $a > l_\varphi$, for low enough temperatures l_φ becomes large, so that $a \ll l_\varphi$ and effectively the sample becomes two-dimensional. The definition of the effective dimensionality of a sample thus depends on the temperature T!
The corrections in smaller dimensions are different and actually more pronounced. Consider the thickness a of a film with $a \ll l_\phi$: the particle will be able to diffuse many times the distance from one wall to another in time τ_ϕ; hence, the probability to find it at any point along the direction of the film normal to the surface is the same.
The general formula for the quantum corrections to the conductivity in $d = 1, 2, 3$ dimensions is given by

$$
\frac{\partial\sigma_d}{\sigma} \simeq -\int_{\tau_{\mathrm{e}}}^{\tau_\varphi}\frac{v_{\mathrm{F}}\lambda_{\mathrm{F}}^2}{(Dt)^{d/2}}\frac{1}{a^{3-d}}\,\mathrm{d}t, \tag{23}
$$

where a is the thickness of a film or the diameter of a wire (depending on the dimensionality d).
For effective dimensionality $d = 1$ and $d = 2$, one obtains:

$$
d=2: \quad \frac{\partial\sigma_2}{\sigma} \simeq -\int_{\tau_{\mathrm{e}}}^{\tau_\varphi}\frac{v_{\mathrm{F}}\lambda_{\mathrm{F}}^2}{Dt}\frac{1}{a}\,\mathrm{d}t \simeq -\frac{1}{k_{\mathrm{F}}l_{\mathrm{e}}}\frac{1}{k_{\mathrm{F}}a}\ln\left(\frac{\tau_\varphi}{\tau_{\mathrm{e}}}\right) \tag{24}
$$

$$
d=1: \quad \frac{\partial\sigma_1}{\sigma} \simeq -\int_{\tau_{\mathrm{e}}}^{\tau_\varphi}\frac{v_{\mathrm{F}}\lambda_{\mathrm{F}}^2}{\sqrt{Dt}}\frac{1}{a^2}\,\mathrm{d}t \simeq -\frac{1}{(k_{\mathrm{F}}a)^2}\left(\frac{l_\varphi}{l_{\mathrm{e}}}-1\right) \tag{25}
$$

Taking into account that $k_{\mathrm{F}}\cdot l \propto \hbar/e^2$ (see equation 14), one can calculate:

$$
d=3: \quad \Delta\sigma_3 \propto -\mathrm{const.} + \left(\frac{e^2}{\hbar}\right) l_\varphi^{-1} \tag{26}
$$

$$d = 2: \quad \Delta\sigma_2 \propto -\left(\frac{e^2}{\hbar}\right) \ln\left(\frac{\tau_\varphi}{\tau_e}\right) \propto -2\left(\frac{e^2}{\hbar}\right) \ln\left(\frac{l_\varphi}{l_e}\right) \tag{27}$$

$$d = 1: \quad \Delta\sigma_1 \propto \text{const.} - \left(\frac{e^2}{\hbar}\right) l_\varphi \tag{28}$$

All the corrections have the same scale $e^2/\hbar \approx 1/4110\,1/\Omega$.
Note: The quantum corrections to the conductivity do not depend on the charge carrier concentration n, although conductivity does ($\sigma \propto n$). Furthermore, the quantum corrections depend less on the scattering times τ than the conductivity does. Therefore, interference corrections become more important with small initial conductivity. Weak localization is thus usually considered as a dirty-metal effect.[6]

Effect on transport properties: experiments To observe weak localization, elastic scattering events thus have to be predominant, so that elastic scattering takes place before the phase information is lost by inelastic or spin-flip scattering. This is realized for a high amount of static defects and low temperatures.
The increased probability of an electron to return to the origin influences transport characteristics in a solid: while the resistance of a classical metal saturates at low temperature T towards ρ_{def}, that one of a dirty metal increases with decreasing temperature T.

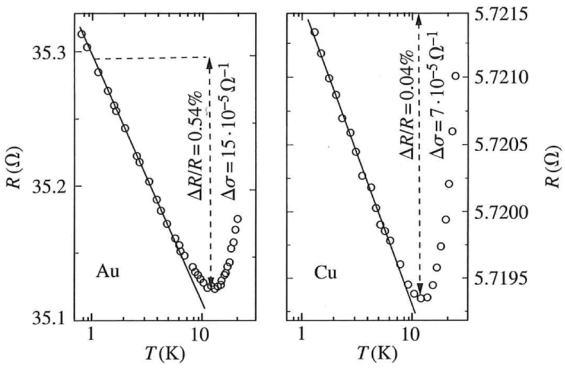


Figure 17: **Weak localization:** Temperature dependence of the resistivity of thin Au [17] and Cu [18]: Both materials feature a logarithmic increase of the resistivity for decreasing temperatures below 10 K. The dashed arrows indicate this increase of resistivity between 10 K and 1 K. Although the relative corrections $\Delta R/R$ of the resistivity differ significantly for these two materials, their corrections $\Delta\sigma$ to the conductivity do much less. Secondary source: [11].

This can be seen in resistivity measurements, as shown in figure 17, where experiments on Au and Cu films are displayed: the temperature dependence of resistivity shows the expected logarithmic increase of the resistivity R with decreasing temperature T between 10 K and 1 K

[6] At low enough temperature, any metal becomes a dirty metal!

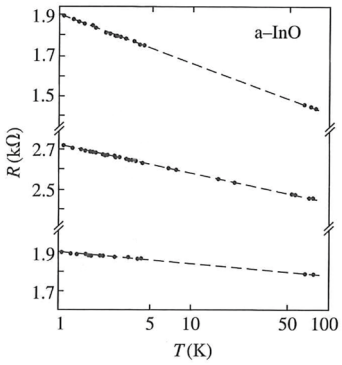


Figure 18: **Weak localization:** Temperature dependence of the resistivity of thin amorphous InO films of different oxygen content [19]: The logarithmic increase of the resistivity with decreasing temperature is already visible at $100\,\mathrm{K}$. The quantum corrections $\Delta\sigma$ to the conductivity are in the range of $(1-2)\cdot 10^{-5}\,\Omega^{-1}$. Secondary source: [11].

predicted by equation 27.[7] Despite the fact that the different materials have different resistivities R and relative corrections $\Delta R/R$ to the resistivity, their corrections $\Delta\sigma$ to the conductivity are almost in the same range: $\Delta\sigma_{\mathrm{Au}} = 15\cdot 10^{-5}\,\Omega^{-1}$ and $\Delta\sigma_{\mathrm{Cu}} = 7\cdot 10^{-5}\,\Omega^{-1}$; this is due to the fact that (26)-(28) to the conductivity depend only weakly on material properties. This also becomes obvious when comparing the correction to the conductivity upon weak localization for amorphous InO (see figure 18): amorphous InO films have a 1000 times higher resistivity and the effects of weak localization is already visible at $T \approx 100\,\mathrm{K}$. However, their quantum corrections $\Delta\sigma$ to the conductivity are in the range of $(1-2)\cdot 10^{5}\,\Omega^{-1}$, which hardly differ from the corrections for Au and Cu.

We have seen that in two dimensions $d = 2$ the quantum corrections $\Delta\sigma$ to the conductivity are proportional to $\ln\left(\frac{\tau_\varphi}{\tau_e}\right)$ (see equation 27). The temperature dependence of $\Delta\sigma$ comes from the temperature dependence of τ_φ: $\tau_\varphi \propto T^{-p}$, so that $\Delta\sigma_{d=2} \propto -\ln(pT)$. We can hence characterize the temperature dependence of τ_φ from low temperature resistance measurements. The dependence of $\Delta\sigma$ on $\ln(T)$ is the hallmark of weak localization. The problem (as we will see later) is that there is a second (very different mechanism) which leads to the same temperature dependence: this is electron-electron interaction.

Weak localization results from electrons interfering with themselves. *Electron-electron interaction*, on the other hand, deals with the interference of waves from different electrons. For example, figure 19 shows the effect of electron-electron interactions on the $\rho(T)$ curve at low-T, which is partially counterbalanced by weak antilocalization in this particular case. The quantum corrections arising from electron-electron interactions will be discussed in detail in section 5.2). Although weak localization and electron-electron interaction look the same from the $\rho(T)$ curve, it's possible to distinguish them with help of a magnetic field, as discussed in the following section.

[7] For small temperatures $T < 10\,\mathrm{K}$ the phase-breaking lengths l_φ become so large that $a \ll l_\varphi$ is fulfilled and the thin Au and Cu films become two-dimensional.

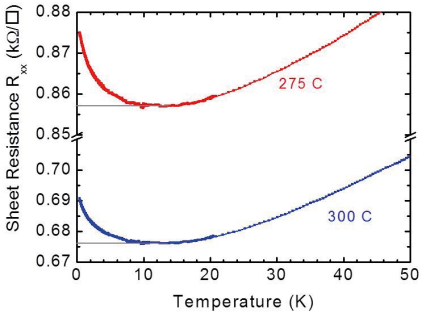


Figure 19: **Sheet resistance at small temperature for a 'dirty metal'** According to equation 10 the sheet resistance R_{xx} of a classical metal is expected to saturate at ρ_{def} at low temperatures. In $GeSb_2Te_4$, however, which is a dirty metal, in fact an increase of the sheet resistance R_{xx} is observed upon cooling. This effects results from an interplay of weak antilocalization, which leads to a decrease of the resistance, and disorder-enhanced electron-electron interactions, which lead to an increase of the resistance. [20]

4.3 Effect of a Magnetic Field on Weak Localization

Analysis of the phenomenon In the previous section the phenomenon of weak localization has been introduced: if elastic scattering is the predominant scattering effect ($k_{\text{F}} \cdot l \gtrsim 1$ and low T), then additional quantum corrections to the conductivity have to be considered in the Boltzmann theory. Indeed, due to the wave-like character of electrons and the phase of an electron not being broken by an inelastic or spin-flip scattering event within τ_φ (which is much larger than the time τ_{e} for elastic scattering), an electron can interfere constructively with itself when being scattered back to the origin within τ_φ: this leads to an increased probability of backscaterring that has to be taken into account in the conductivity by quantum corrections. In this section the effect of a magnetic field, which is crucial to distinguish between weak localization and electron-electron interactions, is discussed.

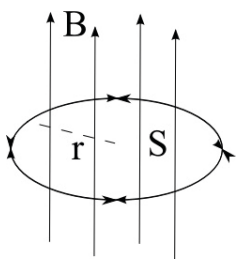


Figure 20: Electron diffusing under the influence of a magnetic field perpendicular to two circular loops with opposite direction that represent two possible electron backscattering paths.

If a sample is placed in a magnetic field B, then the wave function Ψ of a particle passing the loop clockwise and counterclockwise (figure 20) acquires additional phase factors:

$$\begin{aligned}\Psi_1 &\rightarrow \Psi_1 \exp\left(i\frac{e}{\hbar c}\oint \vec{A}\,\mathrm{d}\vec{l}\right) = \Psi_1 \exp\left(\frac{i\pi BS}{\Phi_0}\right)\\ \Psi_2 &\rightarrow \Psi_2 \exp\left(-\frac{i\pi BS}{\Phi_0}\right)\end{aligned}$$

where $\vec{A}$ is the vector potential of the magnetic field; $\Phi_0 = \pi\hbar c/e$ is the quantum of the magnetic flux (in cgs system of units); S is the projection of the loop area on the plane perpendicular to the magnetic field direction; $BS = \phi$ is the magnetic flux.

N.B. Here, the curl theorem has been applied: $\oint \vec{A}\,\mathrm{d}\vec{l} = \iint \mathrm{rot}\vec{A}\,\mathrm{d}\vec{f} = \iint \vec{B}\,\mathrm{d}\vec{f} = BS$.

These additional phase factors of opposite sign for opposite diffusing directions lead to a phase difference $\Delta\varphi$ between the waves of a particle passing along a closed loop clock- and counter-clockwise:

$$\Delta\varphi = 2\pi\frac{BS}{\Phi_0}. \tag{29}$$

Therefore, in the presence of a magnetic field the phases of a particle passing a loop in opposite directions are not coherent any more and the quantum mechanical probability of an electron for returning to the origin (*cf.* equation 20 for the case without any magnetic field) becomes:

$$\begin{aligned}p(\vec{r}=0,t)_{\mathrm{qm},\vec{B}\neq 0} &= |A_1 + A_2|^2\\ &= |A_1^2| + |A_2^2| + 2|A_1||A_2|\cos(\Delta\varphi)\\ &\overset{A_1=A_2}{=} 2A_1^2\left[1+\cos(\Delta\varphi)\right],\end{aligned}$$

where $A_1 = A_2$ are the amplitudes of the wave functions of the electron for passing the loop clock- and counter-clockwise.

For small phase difference $\Delta\varphi \approx 0$ the magnetic field hardly has any influence on weak localization. However, if $\Delta\varphi$ becomes larger, the magnetic field destroys the constructive interference of an electron with itself, so that the probability for a particle to return to a given point and hence the resistivity are reduced. This is the mechanism responsible for the *negative* magneto-resistance.

Now the following question arises: what value B of the magnetic field is sufficient to destroy the constructive self-interference of an electron circulating a loop of area S?

By assuming $S = \pi r_B^2$, where r_B is the *magnetic length* and considering that the effect of the magnetic field is strongest at $cos(\Delta\phi) = -1$, i.e. $\Delta\phi = \pi$, one obtains:

$$B \simeq \frac{\Phi_0}{r_B^2} \tag{30}$$

Using $\Phi_0 = \pi\hbar c/e$:

$$r_B = \left(\frac{\hbar c}{2eB}\right)^{1/2}.$$

The larger the area S, the lower the required B.

N.B. The formulas have been derived in CGS units!

Effect on the conductivity For calculation of the corrections to conductivity due to the presence of a magnetic field the upper integration limit in equation 23 has to be changed: instead of integrating up to τ_φ, one now only takes the diffusion until τ_B, which is the time after which phase information is lost due to the magnetic field, into account. Since for $t > \tau_\phi$ the phase coherence is destroyed by inelastic or spin-flip scattering anyway, $\tau_B < \tau_\phi$ has to be true in order to see the effect of the magnetic field.
τ_B can be estimated from equation 30 with the help of the relation $r_B \propto \sqrt{D\tau_B}$:

$$\tau_B \simeq \frac{\Phi_0}{DB}, \tag{31}$$

which leads to the definition of a critical magnetic field above which the effect is seen: $B_{\text{crit}} = \frac{\Phi_0}{D\tau_\varphi}$.
N.B. The loop represents one possible electron path, but the overall motion is "radial" diffusive. That's why we use r_B as magnetic phase breaking length.
By introducing the *cyclotron frequency*[8] $\Omega = \frac{eB}{m^*c}$ (in cgsunits - see also [6], Panel VIII), one obtains the connection between τ_B and $k_\text{F} l_\text{e}$:

$$\begin{aligned} \tau_B &\simeq \frac{\Phi_0}{DB} && \text{with} \quad D = \frac{l_\text{e} v_\text{F}}{d} \quad \text{and} \quad \Phi_0 = \frac{\pi\hbar c}{e} \\ &= \frac{\pi\hbar c}{eB}\frac{d}{l_\text{e} v_\text{F}} && \text{with} \quad v_\text{F} = \frac{\hbar k_\text{F}}{m^*} \\ &\simeq \Omega^{-1}\left(k_\text{F} l_\text{e}\right)^{-1} \end{aligned}$$

The corrections to conductivity due to the presence of a magnetic field $B > B_{\text{crit}}$ are then given by:

$$0 < \Delta\sigma(B) - \Delta\sigma(0) \approx \begin{cases} 2\frac{e^2}{\hbar}\ln\left(\frac{l_\varphi}{r_B}\right) & (d=2) \\ \frac{e^2}{\hbar}\left[\frac{1}{r_B} - \frac{1}{l_\varphi}\right] & (d=3) \end{cases} \qquad \text{for } l_\text{e} \ll r_B \leq l_\varphi.$$

N.B. l_e doesn't show up in these equations since we subtract the conductivity $\Delta\sigma(0)$ at zero field from $\Delta\sigma(B)$; both, $\Delta\sigma(0)$ and $\Delta\sigma(B)$, depend on l_e.
This phenomenon of magneto conductivity has two interesting features:

- ◇ The effect is visible in classically weak magnetic fields where the conventional magnetic resistance is practically zero.
- ◇ In two dimensions the effect is strongly anisotropic: If the magnetic field is applied in the direction perpendicular to the thin film, there is a pronounced effect, but if it is applied within the plane of the thin film, it is much weaker.[9]

figure 21 displays an example of the destruction of weak localization by the application of a magnetic field. The lower the temperature T, the more pronounced is the decrease of resistivity R with increasing magnetic field B. This effect is particularly intense at low magnetic

[8] The cyclotron frequency is the frequency of an electron moving on a circular path perpendicular to an applied magnetic field.
[9] The reason is that in the latter case the loops and therefore the magnetic flux $\phi = BS$ are much smaller.

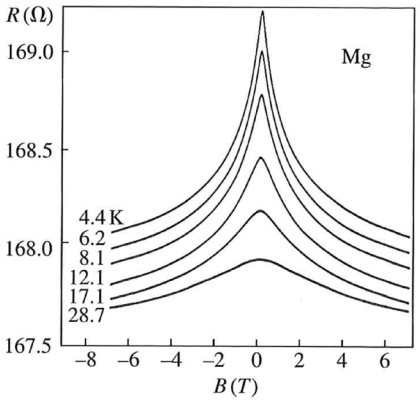


Figure 21: **Magnetoresistivity of a thin Mg film:** With increasing magnetic field B the resistivity R is decreased. This effect is more pronounced at lower temperature T. Primary source: [16], secondary source: [11].

fields. Furthermore, also a hallmark of weak localization is visible: without any magnetic field, $B = 0$, at low temperatures T the resistivity R increases (logarithmically) with decreasing temperature T.
Since electron-electron interaction – just like weak localization – shows a logarithmic increase of the resistivity at very small, decreasing temperatures, but does not show the effect of magnetoresistance, the magnetic field can be used to discriminate whether the former effect comes from electron-electron interaction or from weak localization.

4.4 Summary

So far, in the region of the "yellow traffic light" (*cf.* table 1), *i.e.* $k_{\mathrm{F}} \cdot l \gtrsim 1$, we derived a quantitative theory for perturbations to the Boltzmann equation, so that we can predict the resistance of a metal at low temperature T, which involves quantum corrections. However, two aspects have been ignored so far:

- ◇ Spin flip scattering and spin orbit coupling:
  Spin effects lead to weak antilocalization.

- ◇ Effect of disorder:
  Weak localization and the phenomenon of magnetoresistance result from electrons interacting with themselves. In case of several possible diffusion paths, the partial waves of *one* electron experience different phase shifts, which influence interference at the origin. However, we also have to discuss electron-electron interactions, *i.e.* the interactions of different electrons with one another. These interactions are modified by disorder. This interplay of disorder and electron-electron interactions will be discussed in the following chapter.

5 The Interplay of Disorder and Interactions

Outline

In the previous chapter the influence of the interaction of electrons with themselves on the electronic transport properties of a metal has been discussed. In the regime of $k_\mathrm{F} \cdot l \gtrsim 1$ at low temperatures ($l_e \ll l_\phi$) the electrons in a metal move by diffusive transport and the corresponding waves circulating a closed path clock- and counterclockwise have coherent phases if the length L of the path is smaller than the phase-breaking length $l_\varphi = l_\varphi(T)$: the partial waves of one electron can thus interfere constructively, so that backscattering to the origin is enhanced leading to "weak localization" and the necessity of adding quantum corrections to the conductivity derived according to Boltzmann theory. On the other hand, applying a magnetic field B leads to a phase shift of the partial waves of an electron depending on the orientation of circulation, so that the phase coherence upon diffusive transport is destroyed and the interference is modulated.
In the present chapter, the effect of mutual electron interactions is studied. In particular, we will see:

- ◇ why and how disorder enhances electron-electron interaction, with particular attention to the density of states[10]
- ◇ the existence of a soft Coulomb gap around the Fermi level;
- ◇ the effect of electron-electron interactions on transport;
- ◇ how the density of states can be measured by tunneling spectroscopy.

5.1 Effect of Disorder on the Density of States

Disorder and electron-electron interactions As discussed on page 10, electron-electron interactions can be usually considered a weak effect since the cross section of electron-electron scattering $\Sigma \propto \left(\frac{T}{T_\mathrm{F}}\right)^2 \Sigma_0$ is very small. However, in a highly disordered metal the electrons move by diffusive transport and electrons move away from each other much more slowly, resulting in increase of the mutual interactions. Indeed, the distance r of electrons that were close to each other at $t = 0$ increases linearly with time t in case of ballistic transport (motion without any scattering events):

$$\Delta r_\mathrm{ballistic} \sim v_\mathrm{F} t.$$

On the other hand, the mean distance increases with $\sqrt{t}$ (*cf.* equation 18) in the diffusive regime:

$$\begin{aligned} \Delta r_\mathrm{diffusive} &\sim l_\mathrm{e}\sqrt{\frac{t}{\tau_\mathrm{e}}} \qquad \text{with } l_\mathrm{e} = \tau_\mathrm{e} v_\mathrm{F} \\ &\sim v_\mathrm{F}\sqrt{t\tau_\mathrm{e}}. \end{aligned}$$

Since τ_e is rather small in a highly disordered metal because of the many collisions of electrons with defects of the crystal, the distance between two electrons increases much more slowly upon diffusive transport and the conditions of their interaction are changed.

[10] For a *non-interacting* electron system the density of states $D(E)$ has already been discussed using the Sommerfeld model of a free electron in a box (see equation 2 on page 7).

Electron-electron dephasing time and length In electron-electron interaction, the characteristic dephasing time τ_{ee} depends on the difference ΔE of the initial energies of the electrons. The change of phase $\Delta\varphi(t)$ of one electron with time t is given by the relationship

$$\exp\left[i\varphi(t)\right] \;=\; \exp\left[i\frac{E_i}{\hbar}t\right]$$

where E_i is the initial energy of the electron.
Therefore, the phases of two electrons with energy difference ΔE that have the same phase at $t = 0$ differ by a value of the order of unity after time $\hbar/\Delta E$ has passed. Since only electrons with energies E in the range

$$E_\mathrm{F} - k_\mathrm{B}T \;\lesssim\; E \;\lesssim\; E_\mathrm{F} + k_\mathrm{B}T$$

can reach unoccupied states at temperature T and thus participate in diffusive transport, and their mean energy difference $\overline{\Delta E}$ is proportional to $k_\mathrm{B}T$, the dephasing time τ_{ee} is inversely proportional to T:

$$\tau_{ee} \;\simeq\; \frac{\hbar}{k_\mathrm{B}T}. \tag{32}$$

The longer the dephasing time τ_{ee}, the longer different electrons (which already had the same phase at $t = 0$) keep the same phase. Thus, electron-electron interactions become stronger with decreasing temperature!
The size of the interference region is given by

$$\begin{aligned} l_{ee} &\simeq l_\mathrm{e}\left(\frac{\tau_{ee}}{\tau_\mathrm{e}}\right)^{1/2} \\ &= \frac{l_\mathrm{e}}{\tau_\mathrm{e}}\left(\frac{\hbar\tau_\mathrm{e}}{k_\mathrm{B}T}\right)^{1/2} \qquad \text{with } D = \frac{l_\mathrm{e}^2}{\tau_\mathrm{e} d} \simeq \frac{l_\mathrm{e}^2}{\tau_\mathrm{e}} \\ &\simeq \sqrt{\frac{\hbar D}{k_\mathrm{B}T}}. \end{aligned}$$

Effect on the density of states Electron-electron interactions during diffusion leads to changes of the electron density of states in the vicinity of the Fermi level E_F.
Let's consider the interaction of two electrons with energies $E_\mathrm{F} + E$ and $E_\mathrm{F} - E$. Due to the dephasing time $\tau_{ee} = \frac{\hbar}{E}$ being inversely proportional to E, the effective interaction time is longer the lower the value of $|E|$. It can be proven that, at $T = 0\,\mathrm{K}$ the density of states in the vicinity of the Fermi level E_F depends on the energy E as[11]

$$DOS(T = 0\,\mathrm{K}, E) \sim \text{const.} + (\hbar D)^{-\frac{d}{2}} \begin{cases} \sqrt{|E|} & d = 3 \\ \ln\left(\frac{E\tau_\mathrm{e}}{\hbar}\right) & d = 2 \\ \frac{1}{\sqrt{|E|}} & d = 1 \end{cases} \tag{33}$$

where D is the diffusion coefficient.

[11] Dedicated readers can find the derivation of the density of states in the presence of electron-electron interaction in section 13.4 of [21]. It is also available online: `http://physics.technion.ac.il/~eric/books/chapter13.pdf` (date: 15th July 2013).

Thus, since disorder decreases the diffusion coefficient $D = l_e v_F/d$, the distance between the levels becomes wider and the density of states around the Fermi level E_F decreases. The lower the energy E, the longer the dephasing time $\tau_{ee} \simeq \frac{\hbar}{k_B T} = \frac{\hbar}{E}$, and thus the more pronounced the electron-electron interaction.

5.2 Effect of Electron-Electron Interactions on Transport

Quantum corrections to conductivity While before we had only known about dopants increasing the density of states around the Fermi level E_F, the last section showed that an increase of the disorder in a solid leads to stronger electron-electron interactions, resulting in a decrease of the density of states $g(T \approx 0, E \approx E_F)$ near the Fermi level E_F at low temperatures T due to a decrease of the diffusion coefficient.

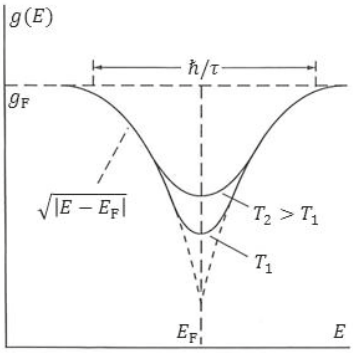


Figure 22: **Density of states in a disordered, three-dimensional sample with significant electron-electron interaction:** The density of states $g(E)$ has its minimum at $E = E_F$. With decreasing temperature T, the electron-electron interaction becomes stronger and the density of states becomes smaller in the vicinity of the Fermi level E_F. [11]

Electron-electron interactions are characterized through the dephasing time τ_{ee}, which is inversely proportional to the temperature T. Therefore, also the density of states depends on temperature($g(E) = g(E,T)$), as shown in figure 22, and so do the quantum corrections to the conductivity .
The corrections to conductivity resulting from electron-electron interactions in a disordered metal can be derived in analogy to the theory of weak localization (see section 4.2). The probability of electrons interfering with other electrons is described by the same integral (*cf.* equation 23), but now the upper integration limit, which in case of weak localization was the phase-breaking length τ_φ, has to be replaced by the dephasing time $\tau_{ee} = \hbar/\Delta E$ with the energy difference ΔE of the interfering electrons since $\tau_{ee} < \tau_\varphi$:

$$\frac{\partial_{ee}\sigma_d}{\sigma} \simeq -\int_{\tau_e}^{\hbar/\Delta E} \frac{v_F \lambda_F^2}{(Dt)^{d/2}} \frac{1}{a^{3-d}}\, \mathrm{d}t, \tag{34}$$

where again: v_F is the Fermi velocity; $\lambda_\mathrm{F} = 2\pi/k_\mathrm{F}$ is the Fermi wavelength; D is the diffusion coefficient $D = \frac{l_\mathrm{e} v_\mathrm{F}}{d}$; d is the dimensionality of the space in which diffusion takes place; a is the thickness of a film or the diameter of a wire (depending on the dimensionality d).
It is not surprising that this similarity in the above expression leads to similar corrections to the conductivity as in the case of weak localization:

$$d = 3: \quad \Delta_\mathrm{ee}\sigma_3 \propto -\mathrm{const.} + \left(\frac{e^2}{\hbar}\right) l_\mathrm{ee}^{-1} \tag{35}$$

$$d = 2: \quad \Delta_\mathrm{ee}\sigma_2 \propto -\left(\frac{e^2}{\hbar}\right) \ln\left(\frac{\tau_\mathrm{ee}}{\tau_\mathrm{e}}\right) \propto -2\left(\frac{e^2}{\hbar}\right) \ln\left(\frac{l_\mathrm{ee}}{l_\mathrm{e}}\right) \tag{36}$$

$$d = 1: \quad \Delta_\mathrm{ee}\sigma_1 \propto \mathrm{const.} - \left(\frac{e^2}{\hbar}\right) l_\mathrm{ee} \tag{37}$$

Electron-electron interactions vs weak localization Compared to the corrections to conductivity due to weak localization (*cf.* equation 26–(28)), only the characteristic parameters τ_φ and l_φ have been replaced by τ_ee and l_ee, while the dependence on the dimensionality stayed the same and still $a \overset{!}{\ll} l_\varphi$. Hence, we need to consider carefully how these two phenomena can be distinguished.
The best indicator for weak localization is the dependence of the resistivity $R(T)$ on the magnetic field B, which is called magnetoresistance (*cf.* section 4.3): The application of a magnetic field changes the phases of the electron's partial wave functions depending on the orientation of their diffusive circulation on a closed path, so that they do not interfere constructively any more. The enhanced probability of backscattering, which is the basis of weak localization, thus is destroyed and a negative magnetoresistance is observed.
On the other hand, the best proof for electron-electron interaction is the density of states at the Fermi energy, that can be probed performing tunneling experiments where a strong decrease of the density of states near the Fermi level upon increasing the disorder and thereby strengthening electron-electron interaction can be observed indirectly, as shown in the next section.

5.3 Determination of the Density of States by Tunneling Spectroscopy

Tunnel junction measurements The energy scheme of experiments to measure the density of states by tunneling is depicted in figure 23; these experiments are important to verify the concepts of disorder and electron-electron interference discussed in section 5.1. The two conducting materials M_1 and M_2 are in contact via an insulating layer I. If the insulating layer is thin enough, (thickness around $(10 - 15)$ Å[12]), electron tunneling is possible and the device is called *tunnel junction*.
In equilibrium, the Fermi levels of the two electrodes M_1 and M_2 coincide: $E_{\mathrm{F},1} = E_{\mathrm{F},2}$. When applying a voltage U to the junction, the potential difference is concentrated within the tunneling gap I, since the resistance of the electrodes is much lower than the resistance of the insulating layer, and a current I flows across the junction:

$$I(U) \propto \int_{-\infty}^{\infty} g_1(E - eU) g_2(E) \left[f_1\left(\frac{E - eU}{k_\mathrm{B}T}\right) - f_2\left(\frac{E}{k_\mathrm{B}T}\right)\right] \mathrm{d}E$$

[12] This requires a reproducible and reliable thickness, so that there is no short-circuit.

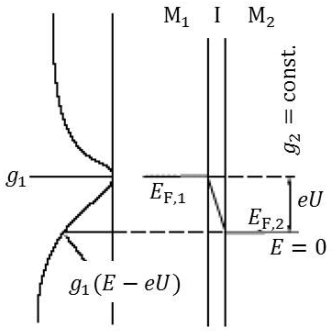


Figure 23: **Energy scheme of current flow through a tunneling junction M_1–I–M_2:** Two conducting materials M_1 and M_2 and an insulating layer I in between form a tunnel junction. In equilibrium, the Fermi levels of both conducting materials are the same: $E_{F,1} = E_{F,2}$. Applying a voltage U to the junction leads to a current through the tunneling gap I. [11]

where g_1 and g_2 are the densities of states of metals M_1 and M_2; f_1 and f_2 are the Fermi distributions of metals M_1 and M_2.
For simplicity, we now assume that one electrode is a conventional metal with g_2 =const. and that the temperature T is very low, so that the Fermi distribution becomes a step function. The current I across the junction then is given by:

$$I(U) \propto g_2 \int_0^{eU} g_1(E)\mathrm{d}E.$$

When adding an AC voltage $U_{\mathrm{AC}} = U_\omega \sin(\omega t)$ to the DC voltage $U_{\mathrm{DC}} = U$, the current will feature an additional, alternating term I_ω with frequency ω, which is proportional to the derivative $\mathrm{d}I/\mathrm{d}U$:

$$\begin{aligned} I(U + U_\omega \sin(\omega t)) &= I(U) + \frac{\mathrm{d}I}{\mathrm{d}U} U_\omega \sin(\omega t) \\ I_\omega &\propto \frac{\mathrm{d}I}{\mathrm{d}U} \propto g_1(eU). \end{aligned} \tag{38}$$

This correlation enables the determination of the density of states with the help of tunneling measurements.

Soft Coulomb gap Figure 24 shows an important example. Here a tunneling junction consisting of Pb-SiO_2-Si:B is studied. At these low temperatures of $T = 1.15$ K Pb is superconducting and hence the superconducting gap appears in the density of states. If a magnetic field $H \neq 0$ is applied, then the superconductivity is destroyed and the measured density of states changes. In this case, a parabolic gap due to disorder and/or electron correlations and called *soft Coulomb gap* appears around E_F; the gap is called "soft" because the density of states only vanishes exactly at the Fermi energy E_F and increases with $|E - E_F|^2$.

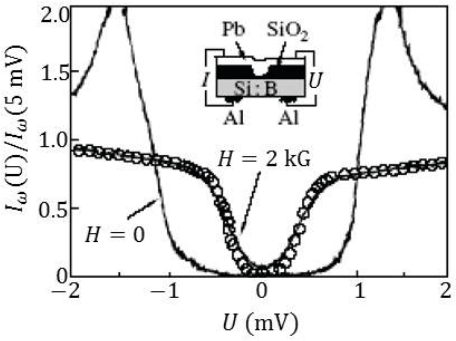


Figure 24: **Normalized current as a function of voltage bias in the tunnel junction Pb–SiO_2–Si:B:** The measurements have been performed at $T = 1.15$ K, so that Pb is superconducting. In zero magnetic field $H = 0$ there is a superconducting gap of the current between $U = -1$ mV and $U = +1$ mV, since there are no states around the Fermi edge. The application of a magnetic field $H = 2$ kG destroys superconductivity of Pb. However, there still remains a (smaller) gap which comes from the disorder and/or electron correlation effects in Si:B. Primary source: [22], secondary source: [11].

N.B. The Coulomb gap is a phenomenon arising of the density of states due to electron correlations in an insulator, while the density of states derived in equation 33 is changed due to electron correlations in a metal.

Another example Figure 25 shows another interesting example obtained for a tunneling junction of $Ge_{1-x}Au_x$-Al_2O_3-Al, where Al is the counter electrode and $Ge_{1-x}Au_x$ is studied as a function of the Au concentration x. For high x the system is metallic and we observe a minimum of $\mathrm{d}I/\mathrm{d}U$ at the Fermi level $E_\mathrm{F} \equiv 0$ in accordance with the theoretical considerations in equation 33 and (39); this is due to the interaction of diffusing electrons. We have discussed this effect within the framework of perturbation theory, *i.e.* under the assumption that corrections to the density of states $g(E)$ are small. With decreasing Au concentration x, the dip of $\mathrm{d}I/\mathrm{d}U$ at $E_\mathrm{F} \equiv 0$ becomes more significant and the alloy approaches its metal-insulator transition; at $x = 0.08$, $\mathrm{d}I/\mathrm{d}U$ and thus $g(E)$ go to zero at $E_\mathrm{F} \equiv 0$ and can be described by a parabola in the vicinity. This indicates formation of a Coulomb gap. In conclusion, not only the presence of the minimum of the density of states $g(E \approx E_\mathrm{F})$ in a dirty metal due to electron-electron interactions

5.4 Summary

In chapters 4 and 5 we have discussed the corrections to the Boltzmann equation at low temperatures and upon $k_\mathrm{F} \cdot l \gtrsim 1$ (*cf.* table 1). Perturbation theory was necessary to describe the effect of interference of electrons with themselves, *i.e.* weak localization, and of electron-electron interaction, which in case of disorder can take place. This led to quantum corrections to the conductivity of the metallic state, which are negative and increase with decreasing temperature, *i.e.* these phenomena can lead to $\frac{\mathrm{d}\rho}{\mathrm{d}T} < 0$ for low enough temperatures. In the past, this was seen

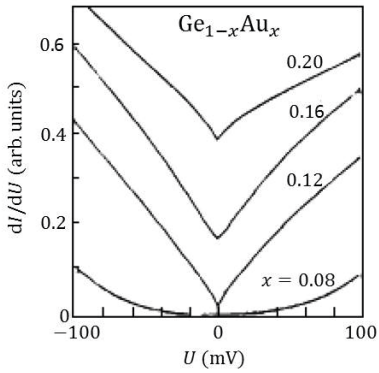


Figure 25: **dI/dU as a function of voltage bias U in the tunnel junction $Ge_{1-x}Au_x$–Al_2O_3–Al:** At high Au concentrations x the alloy is metallic. This dirty metal shows a minimum of dI/dU around the Fermi level E_F. With decreasing x, the dip becomes more significant and the system performs an MIT. The strongly disordered insulator at $x = 0.08$ shows the existence of a Coulomb gap. (The Coulomb gap has not been discussed in the lecture. Dedicated students can get information in section 3.3 in [11].) Primary source: [23], secondary source: [11].

as an indicator for the insulating state. However, the systems we discussed here are still clearly metallic, *i.e.* have a finite density of states at the Fermi energy $D(E_F) \neq 0$ and a non-vanishing charge carrier mobility even at $T = 0\,\mathrm{K}$. In the following chapter the metal-insulator transition will be discussed.

6 Metal-Insulator Transition

Outline

In the last two chapters, transport phenomena in the regime of $k_F \cdot l \gtrsim 1$ (*cf.* table 1 on page 15) at low temperatures have been discussed: phenomena such as weak localization and electron-electron interactions led to the need of quantum corrections to the conductivity, but the systems were still clearly metallic. In this chapter we will now consider what happens if a metal becomes insulating: such a *metal-insulator transition* (MIT) takes place when $k_F \cdot l \approx 1$.

6.1 MIT pathways

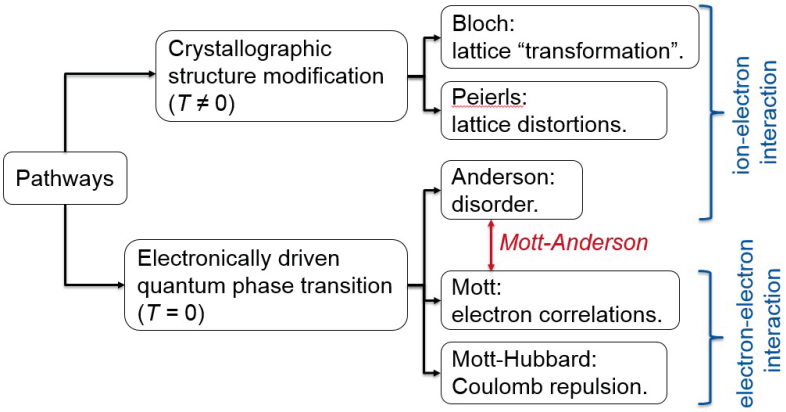


Figure 26: **Scheme of the possible metal-insulator transitions for a material:** MIT pathways can be separated into ones involving changes in the crystallographic structure and ones involving changes in the electronic structure; furthermore, it's possible to distinguish among MITs related to ion-electron interactions and MITs related to electron-electron interactions.

Several pathways for a material to experience a MIT have been identified by scientists in the years; figure 26 sketches them.
The interactions leading to the transition can be ion-electron or electron-electron type and can "tuned" by changing the crystallographic and/or the electronic structure of the material through proper methods. In particular, transitions due to changes in the crystallographic structure are temperature dependent and occur at $T > 0$ K, while transitions due to changes in the electronic structure are 0 K phenomena.
In the following of this section, those mechanisms are investigated.

Change of crystal structure A change of the crystal structure of the material can result in a very different band structure, which can lead to a transition from a metal to an insulator or vice versa.
An example of MIT driven by a change in the crystallographic structure has been already discussed in section 2.1, where data for vanadium dioxide VO_2 have been presented.

Another material that experiences a similar phenomenon is tin, that undergoes a phase transition from a metallic state (white tin) above $18°$ C, with tetragonal crystalline structure, to an insulating state (gray tin) below this temperature, with diamond cubic crystalline structure (this is demonstrated in a video [24]).

> *Anecdote: It has been argued that Napoleon lost the Russian war since his soldiers had buttons made of tin. However this statement is questionable for two reasons: Firstly, graves with French soldiers show no such defective buttons, and secondly, the transformation to grey tin at low temperature should take 18 months, but the battle only lasted for 9 months.*

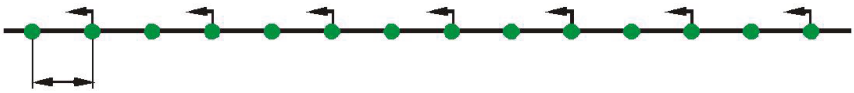


Figure 27: **Peierls distortion:** atoms move closer to one neighbor and further away from the other.

Peierls transition Peierls theorem states that a one-dimensional equally spaced chain with period a of ions with one electron each is unstable. This results in a distortion of the periodic lattice called *dimerization* (figure 27) where atoms move closer to one neighbor and further away from the other, doubling the periodicity to $2a$.
Indeed, the energy gap can be found at k values multiple of $\frac{\pi}{a}$ for the non-distorted crystal and the band is half-filled up to $k = \pm\frac{2\pi}{a}$, resulting in a metallic system.
After dimerization, the doubled periodicity results in the opening of a gap at k values multiple of $\frac{\pi}{2a}$, slightly reducing the electron energy and making the system insulating at low T.

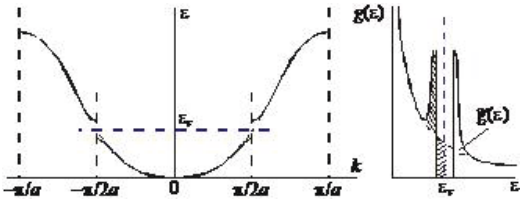


Figure 28: **Band structure and density of states modifications induced by Peierls distortion:** dimerization leads to gap opening in correspondence of the highest occupied state, resulting in a slightly reduction of the energy of the system.

The band diagram and density of states modifications are summarized in figure 28.

Mott-Hubbard transition Before discussing Mott-Hubbard insulators, it's important to introduce the *Hubbard model*.
Hubbard model has been introduced to account for strong electron-electron Coulomb repulsion and consists of the following Hamiltonian:

$$H = -\sum_{ij} t_{ij} c^{\dagger}_{i\sigma} c_{i\sigma} + U \sum_{i} n_{i\uparrow} n_{i\downarrow} \tag{39}$$

where i, j labels the lattice sites; t is the kinetic term from the tight-binding model allowing for tunneling of electrons between sites of the lattice; U is a potential term consisting of an on-site interaction and representing the Coulomb energy necessary to bring a second electron into the same state of another one.

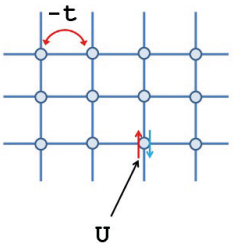


Figure 29: **Hubbard model:** the crystal is represented as a set of lattice sites where electrons can jump from one site to another thanks to their kinetic energy t and repell each other through a Coulomb potential U.

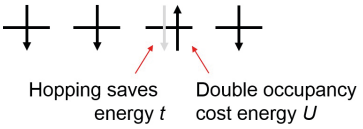


Figure 30: **Mott-Hubbard insulators:** Coulomb repulsion among electrons prevents their transport through the material.

The Hubbard model predicts the existence of Mott-Hubbard insulators, where Coulomb repulsion among electrons prevents electron "jumping" from one site to another destroying charge transport (figure 30).

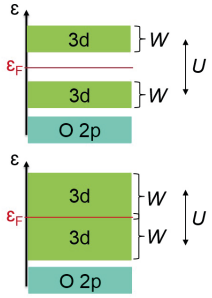


Figure 31: **Mott-Hubbard transition in metal oxides:** an energy gap for $U > W$ and the material is insulating, while the material is metallic for $U < W$. The band width W is a measure of the kinetic energy of the electrons.

By considering transition metal oxides as an example and by defining W as the bandwidth of the outermost electronic state, which is an estimate for the coupling of an electron with its neighboring atoms, it's possible to state that an energy gap is created because of strong electron-electron interaction effects for $U > W$ and the material is insulating, while Coulomb repulsion is not large enough to create an energy gap for $U < W$, under these conditions, the material is metallic and the one-electron approximation holds (figure 31). On the other hand, if $U > W$, a Mott insulator has been created.
N.B. The parameter W will be used again in the following, but it will have a different meaning then!

Anderson transition A MIT induced by disorder without any electron-electron interaction is called *Anderson transition*. The corresponding insulating state is called an *Anderson insulator*. Increasing disorder leads to an increase of scattering events that affect the motion of the charge carriers until the material becomes insulating for $k_F \cdot l \approx 1$.

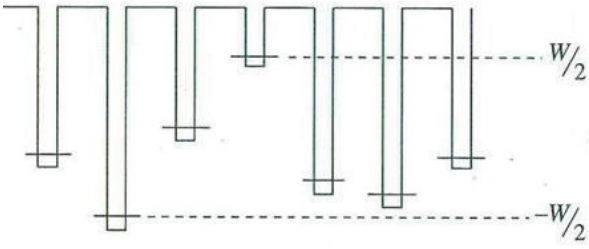


Figure 32: **Kronig-Penney model involving disorder:** Periodically arranged wells describe a one-dimensional lattice. The disorder is modeled by different depths of these quantum wells, where the distribution of these depths is one of the characteristic quantities. The width of this distribution is described by the parameter W. (Adapted from [11].)

A model to study the effect of disorder on transport properties can be defined by using a Kronig-Penney-like model where the potential distribution arising from the ion cores of the crystal is assumed to consist of rectangular potential wells of different depths (see figure 32)[13]. The width W of the distribution characterizes the degree of disorder.[14], while the transfer integral J describes the probability for an electron to 'jump' from one site to the next:

$$J \;=\; \int \psi_1^* \hat{H} \psi_2 \, \mathrm{d}^3 r \;\propto\; \exp\left(-\frac{r_{12}}{a_\mathrm{B}}\right),$$

where ψ_1 and ψ_2 respectively describe the initial and final sites of such jumps; $\hat{H}$ is the Hamiltonian; r_{12} is the distance between two neighboring wells; a_B is the Bohr radius.

The Hamiltonian $\hat{H}$ acts on ψ_2 and the new state is projected onto ψ_1, *i.e.* the probability for the electron in state 2 to move to state 1 is calculated.

In a perfect crystal, $W = 0$, *i.e.* the depths of the potential wells of all the ion cores is equal. If furthermore the transfer integral $J \neq 0$, so that electrons can 'jump' to other sites, the crystal features a finite electron mobility and hence is a metal.

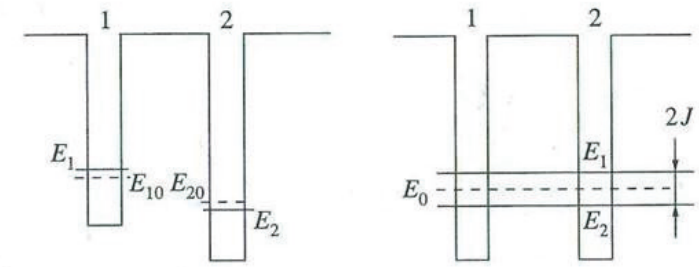
If however $J = 0$, so that every electron is 'captured' in its own well, or W is large, so that the depths of the potential wells differ significantly, the ratio $\frac{J}{W}$ is small and the crystal is an insulator.

The significance of the ratio $\frac{J}{W}$ can be explained by simple quantum mechanics. Consider two wells of different depths with unperturbed levels E_{10} and E_{20} and wave functions φ_1, φ_2 (*cf.* figure 33a).[15] The finite overlap leads to small corrections to the wave functions ψ_1 and ψ_2:

[13] For simplicity, the quantum wells here are assumed to be one-dimensional.

[14] On the previous page the efficiency of the hopping process of the electrons was characterized by W. Now, instead W measures the disorder!

[15] While the wave functions ψ_1 and ψ_2 shall describe the electrons' states taking neighboring states into account, φ_1 and φ_2 with corresponding levels E_{10} and E_{20} are the unperturbed wave functions considered in isolation.



a b

Figure 33: (a) Two quantum wells of different depths with levels E_{10} and E_{20}. (b) Splitting of the level E_0 in case of equivalent levels $E_{10} = E_{20} = E_0$ due to the overlap of wave functions into $E_{1,0} \simeq E_0 \pm J$. [11]

$$\begin{aligned} \psi_1 &= c_1\varphi_1 + c_2\varphi_2 & c_2 &= \frac{J}{E_{10} - E_{20}} \\ \psi_2 &= -c_2\varphi_1 + c_1\varphi_2, & c_1 &\approx 1. \end{aligned}$$

If $c_2 \ll c_1 \approx 1$, the correction is very small and all the electrons are predominantly located in their own wells. In case of equivalent wells, *i.e.* wells of the same depths $E_{10} = E_{10} = E_0$, the level E_0 is split into $E_{1,2} \simeq E_0 \pm J$ due to the overlap of the wave functions (see figure 33b). The resulting wave functions

$$\psi_{1,2} = \frac{1}{\sqrt{2}}(\varphi_1 \pm \varphi_2)$$

are smeared over both wells. The cases of different and equivalent energy levels do not only differ regarding their wave functions, but also regarding their magnitude of energy shifts (*cf.* figure 33). This shall be quantified in the following.
In the model of a one-dimensional rectangular wells as discussed above, in case of wells with different depths, each electron located in one well is perturbated by the electron in the other well, respectively. The shift $\Delta_1 E$ of the level E_{10} is of the order of

$$\Delta_1 E = E_1 - E_{10} \simeq \int \varphi_1^* \hat{H}_2 \varphi_1 \, \mathrm{d}^3 r \propto \exp\left(-\frac{2r_{1,2}}{a_\mathrm{B}}\right)$$

due to the undisturbed wave function φ_1 behaving like $\exp(-r_{12}/a_\mathrm{B})$. In fact, the energy shift $\Delta_1 E$ contains the square of the factor $\exp\left(-\frac{r_{12}}{a_\mathrm{B}}\right)$ of the transfer integral J. In case of equivalent wells the energy shift ΔE is given by

$$\Delta E \sim J \Rightarrow \Delta E \propto \exp(-r_{12}/a_\mathrm{B}).$$

If the difference of unperturbed wave functions $|E_{10} - E_{20}|$ of wells 1 and 2 becomes smaller than the transfer integral J, the wells cannot be treated as "resonant" any more, but the electrons can spread over the wells and their wave functions become delocalized. The ratio J/W thus is the fraction of resonant wells, so that the critical value $\left(\frac{J}{W}\right)_{\text{crit}}$ may be interpreted as a "percolation threshold". The criterion for an Anderson insulator thus is found to be

$$\frac{J}{W} < \left(\frac{J}{W}\right)_{\text{crit}}$$

with the wave functions at the Fermi level being localized.

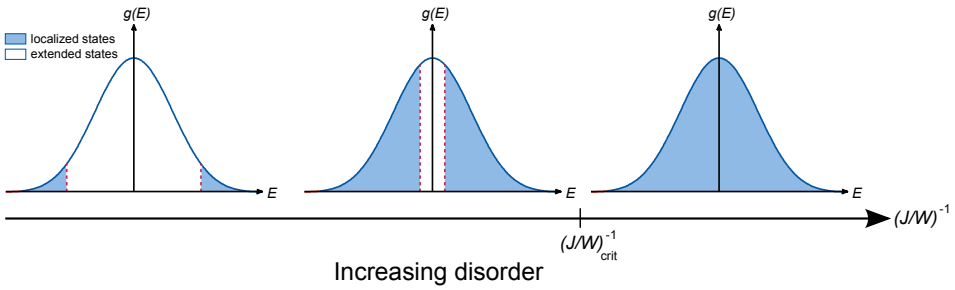


Figure 34: **Anderson transition:** With increasing disorder, *i.e.* with decreasing value for J/W, delocalized states (blue) around the Fermi level E_{F} occupied by electrons become localized, so that a transition from a metal to an Anderson insulator is performed at $(J/W)^{-1} < [(J/W)_{\text{crit}}]^{-1}$.

Such an Anderson insulator differs conceptionally from a band insulator: While in a band insulator the Fermi level is located within the forbidden gap where there are no states, in an Anderson insulator there is a finite number of states near the Fermi level, which are occupied by *localized* electrons (see figure 34).

Mott transition In the previous subsection, it has been shown that increasing disorder reduces electron wave functions overlap between adjacent atoms and electrons become localized within their respective potential wells. This MIT is called Anderson transition.
Another mechanism for MIT, called *Mott transition* after Sir Francis Mott, establishes the transition from an insulator to a metal by an increase of charge carrier concentration induced by doping that leads to overlap of single electron wave functions, resulting in delocalized states where electrons can move from ion to ion and subsequent strong enhancement of electron transport.
The Mott MIT is displayed and compared with the Anderson MIT in figure 35.
Let's now look at the Mott transition in more detail. Two important parameters are the *mean distance between electrons* $n^{-\frac{1}{3}}$ and the *Bohr radius* a_B, representing the most probable distance between the electron and the positively charged core in a dopant atom and defined as:

$$a_B = \frac{4\pi\epsilon_S\hbar^2}{m^* e^2}, \tag{40}$$

where ϵ_S is the dielectric constant of the bulk material.

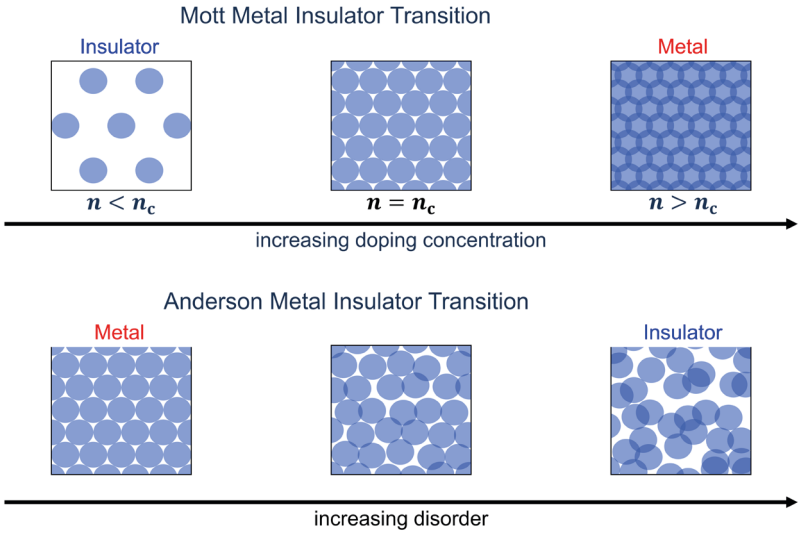


Figure 35: **Metal-insulator transition theories:** The Mott transition, which is displayed on top, explains the transition from the insulating to the metallic state by the electrons becoming delocalized above a critical charge carrier concentration n_{crit} due to the overlap of their wave functions. In the Anderson transition, an increase of disorder leads to an increase of defects, which act as scatter centers and localizes the electrons: The metal transforms into an insulator. In real metal-insulator transitions, both aspects have to be taken into account.

Increasing carrier concentration corresponds to an increase of the screening of the potential generated by atom nuclei. As result, electrons far from a specific nucleus will "feel" the Coulomb interaction with the nucleus less. The overall effect is the reduction of the amount of bound states for electrons in the most external shell until their states become extended and the material turns into a metal thanks to significant overlap of the electron wave functions of adjacent atoms. The last statement can be easily understood by modeling an atom as quantum well: reducing the depth corresponds to reduce the amount of bound states. This is modeled by means of the *screening radius* r_e, that represents the minimum distance from the nucleus at which the electron doesn't feel the potential generated by the nucleus itself:

$$r_e = \left(\frac{4\pi m^* e^2 n^{\frac{1}{3}}}{\epsilon_S \hbar^2} \right)^{-\frac{1}{2}}. \tag{41}$$

Since the electron is most likely to be found at distance a_B from the dopant core, the transition point can be fixed at $a_B = r_e$, while the material is insulating for $r_e > a_B$ and metallic for $r_e < a_B$.
By combining our equations, we can determine a relation between Bohr radius and critical density which leads to the definition of the *Mott line* (figure 36) [Edwards and Sienko, "Universality aspects of the metal-nonmetal transition in condensed media", *Phys. Rev. B*, **17**, 6 (1978)]:

$$a_B n^{\frac{1}{3}} = 0.25. \tag{42}$$

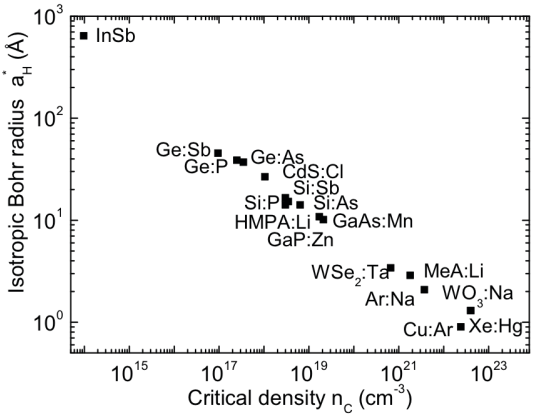


Figure 36: **Mott line:** the figure displays the Mott line, which defines the relation between the Bohr radius and the critical carrier density at the metal-insulator transition.

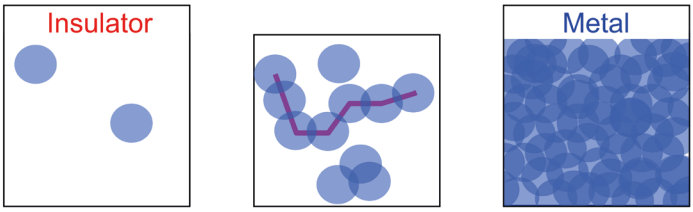


Figure 37: **Real metal-insulator transition:** In real metal-insulator transitions, both aspects, the transition by an increase of disorder as well as the transition by the localization of electrons due to the decrease of overlap of electron wave functions have to be considered. Therefore, real electronically-driven metal-insulator transitions may be called Mott-Anderson transitions.

Mott vs Anderson transition Although Anderson's transition and Mott's transition as displayed in figure 35 appears to be very different phenomena, they are really difficult to be separated in real materials.
There are several reasons for this. As can be seen in figure 37 increasing the concentration of dopants in a semiconductor also simultaneously changes the disorder in the sample. Yet, if we only change the disorder in a material, the electrons will have a more diffusive motion and hence stronger electron-electron interactions will occur. Increasing disorder hence leads to more pronounced electron-electron interactions. Due to this difficulty to distinguish the effects of disorder and electron-electron interactions, we now often talk of an *Mott-Anderson transition*.

References

[1] P. Merkelbach. I. Institute of Physics (IA), RWTH Aachen University. Cited on page 3.

[2] W. H. Rosevear and W. Paul. Hall Effect in VO_2 near the Semiconductor-to-Metal Transition. *Physical Review B*, 7:2109–2111, March 1973. Cited on page 5.

[3] T. D. Manning and I. P. Parkin. Atmospheric pressure chemical vapour deposition of tungsten doped vanadium (IV) oxide from $VOCl_3$, water and WCl_6. *Journal of Materials Chemistry*, 14(16):2554–2559, 2004. Cited on page 4.

[4] The official website of the Nobel Prize. The Nobel Prize in Physics 1977. `http://www.nobelprize.org/nobel_prizes/physics/laureates/1977`. Accessed 12th April 2013. Cited on page 4.

[5] The official website of the Nobel Prize. The Nobel Prize in Physics 1967. `http://www.nobelprize.org/nobel_prizes/physics/laureates/1967`. Accessed 15th April 2013. Cited on page 6.

[6] Harald Ibach and Hans Lüth. *Festkörperphysik: Einführung in die Grundlagen*. Springer, 2008. Cited on pages 7, 9, 10, 11, 13, 16, 18, and 27.

[7] Neil W. Ashcroft and N. David Mermin. *Solid State Physics*. Saunders College, Harcourt College, 1979. Cited on pages 7 and 9.

[8] D. K. C. MacDonald and K. Mendelssohn. Resistivity of pure metals at low temperatures I. The alkali metals. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 202(1068):103–126, 1950. Cited on page 11.

[9] J. O. Linde. Elektrische Eigenschaften verdünnter Mischkristallegierungen. II. Widerstand von Silberlegierungen. *Annalen der Physik*, 406(4):353–366, 1932. Cited on page 11.

[10] J. H. Mooij. Electrical Conduction in Concentrated Disordered Transition Metal Alloys. *Physica Status Solidi (a)*, 17(2):521–530, 1973. Cited on pages 12, 13, and 14.

[11] V. F. Gantmakher and L. I. Man. *Electrons and Disorder in Solids*. Clarendon Press – Oxford University Press, 2005. Cited on pages 12, 13, 14, 23, 24, 28, 31, 33, 34, 35, 39, and 40.

[12] Z. Fisk and G. W. Webb. Saturation of the High-Temperature Normal-State Electrical Resistivity of Superconductors. *Physical Review Letters*, 36(18):1084–1086, 1976. Cited on pages 12 and 13.

[13] A. L. Efros and B.I. Shklovskii. Electronic properties of doped semiconductors. *Springer Series in Solid-State Sciences, Springer, Berlin*, 1984. Cited on page 16.

[14] H. Volker. *Disorder and electrical transport in phase-change materials*. PhD thesis, RWTH Aachen University, 2013. Cited on page 16.

[15] T. Siegrist, P. Jost, H. Volker, M. Woda, P. Merkelbach, C. Schlockermann, and M. Wuttig. Disorder-induced Localization in Crystalline Phase-Change Materials. *Nature Materials*, 10(3):202–208, 2011. Cited on pages 16 and 17.

[16] G. Bergmann. Weak Localization in Thin Films: A Time-of-Flight Experiment with Conduction Electrons. *Physics Reports*, 107(1):1–58, 1984. Cited on pages 19, 21, and 28.

[17] S. I. Dorozhkin and V. T. Dolgopolov. Nonlinearity of the voltage–current characteristics of thin gold films. *JETP Letters*, 36(1), 1982. Cited on page 23.

[18] L. Van den Dries, C. Van Haesendonck, Y. Bruynseraede, and G. Deutscher. Two-Dimensional Localization in Thin Copper Films. *Physical Review Letters*, 46:565–568, Feb 1981. Cited on page 23.

[19] Z. Ovadyahu and Y. Imry. Magnetoconductive Effects in an Effectively two-dimensional System. Weak Anderson Localization. *Physical Review B: Condensed Matter;(United States)*, 24(12), 1981. Cited on page 24.

[20] N. P. Breznay, H. Volker, A. Palevski, R. Mazzarello, A. Kapitulnik, and M. Wuttig. Weak antilocalization and disorder-enhanced electron interactions in annealed films of the phase-change compound $GeSb_2Te_4$. *Physical Review B*, 86(20):205302, 2012. Cited on page 25.

[21] Eric Akkermans. *Mesoscopic physics of electrons and photons*, chapter 13 – Interactions and Diffusion, section 13.4 – Density of States Anomaly. Cambridge University Press, 2007. Cited on page 30.

[22] J. G. Massey and M. Lee. Direct Observation of the Coulomb Correlation Gap in a Nonmetallic Semiconductor, Si:B. *Physical Review Letters*, 75(23):4266–4269, 1995. Cited on page 34.

[23] W. L. McMillan and J. Mochel. Electron Tunneling Experiments on Amorphous $Ge_{1-x}Au_x$. *Physical Review Letters*, 46(8):556–557, 1981. Cited on page 35.

[24] Youtube. Grey tin (tin pest) time-lapse video. `http://www.youtube.com/watch?v=FUoVEmHuykM&list=UUuBFcUuKwlKsws-21CIBELg&index=44`. Accessed 17th July 2013. Cited on page 37.

Recommended Literature

(1) Modern Problems in Condensed Matter Physics.
Volume 10: Electron-Electron Interactions in Disordered Systems.
A. L. Efros and M. Pollak.
North-Holland Physics Publishing 1985.
Chapter 1–3.

(2) Disordered electronic systems.
P. A. Lee and T. V. Ramakrishnan.
Reviews of Modern Physics, Volume 57, No. 2, 287–337, 1985.
(The classic, cited 4091 times as of December 26$^{\text{th}}$ 2015.)

A 8 Magnetism and Spin-Polarized Transport

Daniel E. Bürgler
Peter Grünberg Institut, Elektronische Eigenschaften (PGI-6)
Forschungszentrum Jülich

Contents

1 Introduction **3**

2 Magnetic moments in solids **4**

3 Interaction of magnetic moments **6**
3.1 Dipole interaction 6
3.2 Direct exchange 7
3.3 Heisenberg spin Hamiltonian 8
3.4 Indirect exchange 9
3.5 Superexchange 10
3.6 Dzyaloshinskii-Moriya interaction 11
3.7 Itinerant exchange 12

4 Band magnetism (Itinerant magnetism) **12**
4.1 Ferromagnetic 3d-metals: Fe, Co, Ni 14
4.2 Antiferromagnetic 3d-metals: *e.g.* Mn and Cr 15
4.3 Ferromagnetic 3d-alloys 16

5 Spin-orbit coupling **16**

6 Collective magnetism **19**
6.1 Ferromagnetic order 20
6.2 Ferrimagnetic order 22
6.3 Antiferromagnetic order 23

7 Magnetic anisotropy **24**
7.1 Phenomenology of magnetic anisotropy 25
7.2 Physical origin of magnetic anisotropy 26
7.2.1 Shape anisotropy 26

7.2.2 Magnetocrystalline anisotropy 27
7.2.3 Exchange anisotropy (Exchange biasing) 29
7.3 Superparamagnetism 31

8 Magnetic domains **32**
8.1 Origin of magnetic domains 32
8.2 Domain walls 33

9 Electrical transport in magnetic metals **34**
9.1 Boltzmann equation and relaxation time approximation 35
9.2 Normal and spin-disorder magnetoresistance 36
9.3 Spin polarization and spin accumulation 37

10 Anisotropic magnetoresistance (AMR) **38**
10.1 Phenomenological description 38
10.2 Microscopic picture: Scattering into spin-orbit-coupled states 39

11 Giant magnetoresistance (GMR) **41**
11.1 Phenomenological description 41
11.2 Microscopic picture: Spin-dependent scattering 43

12 Tunneling Magnetoresistance (TMR) **45**
12.1 Phenomenological description 45
12.2 Microscopic picture: Spin-dependent tunneling 45
12.3 Beyond Jullière's model 47

13 Spin-transfer torque (STT) **48**
13.1 Phenomenological description of STT 49
13.2 Physical picture of STT: Absorption of the transverse spin current component . 50
13.3 Extended Gilbert equation and spin-torque oscillators 53
13.4 Current-driven domain wall motion 54

Lecture Notes of the 47^{th} IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

Ever since the discovery of magnetism in the form of "attractive" ores more than 2000 years ago, "invisible" magnetic forces, attractive or repelling in nature depending on how magnets are approached to each other, have fascinated mankind. Magnetism and magnetic phenomena are strongly correlated to specific materials. Pieces of the naturally-occurring mineral magnetite (Fe_3O_4) called "lodestones" were the first magnets known to mankind. Lodestone found in *Magnesia* in Anatolia is most likely the origin of the term "Magnetism".

Throughout its history magnetism is closely related to technological applications. The oldest known application of a magnetic material is a compass made from magnetite introduced by Chinese scholars about 2500 years ago. But only in the 16^{th} century the first scientific study on magnetism called *De Magnete* [1] was published by William Gilbert. Remarkably, in *De Magnete* Gilbert also studied static electricity produced by rubbing amber (*elektron* in Greek) and was the first to clearly distinguish between magnetism and static electricity. The link between classical *electrodynamics* and magnetic phenomena is established in the Maxwell equations that were published by James Clerk Maxwell in the 1860s [2]. Maxwell's equations describe electrical and magnetic phenomena in terms of electric and magnetic fields and how these fields are generated and altered by each other and by electrical charges and currents. These concepts form the foundation for most modern applications of magnetism, which range from permanent magnets at everybody's fridge and installed in huge quantities in wind turbines, omnipresent electro-motors, transformers, and sensors, *etc*. to magnetic data storage. The latter has significantly contributed to the fast-paced, exponential progress in information technology and its applications in almost all aspects of human life from the mid 1950s to today and still is one of the cornerstones of recent developments such as Big Data, Cloud Computing, and Internet of Things.

The discoveries of interlayer exchange coupling and the giant magnetoresistance effect (GMR) in the 1980s, which became the cornerstones of the new field **spintronics**, brought the fusion between magnetism and electronics to a next, quantum-mechanical level. Spintronics comprises all previously neglected effects of the spin on electrical transport and is nowadays a very active scientific research field. From a technological point of view, spintronics represents a new paradigm of electronics based on the electron spin in addition to or even instead of the electron charge offering the advantages of non-volatility, increased integration density, higher processing speed, and enhanced energy efficiency.

From a fundamental point of view the understanding of the origin of magnetism and the phenomena of spintronics are based on quantum mechanics, in particular on the concept of spin that arises from the relativistic treatment of an electron in an electromagnetic field in the framework of the Dirac equation. The fascinating, yet complicated quantum phenomena resulting from the interaction of many electrons (spins) in condensed matter are not completely understood, nor exploited. Therefore, magnetism was and in conjunction with spintronics still is a major driving force for progress in solid-state physics. Complications not only arise from the large number of interaction particles, but also from a large variety of length, energy, and time scales involved. Atomic-scale exchange interactions with coupling constants in the eV range compete with long-range dipole-dipole interaction that depends on the size and geometry of the magnetic object, and with Zeeman and anisotropy energies hardly exceeding the μeV range. Ultra-fast demagnetization occurs on the femtosecond time-scale, whereas data storage applications requires data retention times of 10 years. In addition, finite temperature and collective as well as local excitations have to be considered. A microscopic theory describing all these circumstances

is still missing, which forces us to employ a hierarchy of partly phenomenological theories and models ranging from *ab-initio* quantum-mechanical treatment to micromagnetic modeling and continuum theory.
The goal of this lecture is to introduce the fundamentals and concepts of magnetism and spin-transport that provide the basis of magnetic data storage and spintronics. Hence, I will focus on magnetism in solids and spin-polarized transport in magnetic multilayer structures.

2 Magnetic moments in solids

The classical idea following from Maxwell's equations that a stationary ring current is equivalent to a magnetic moment also holds for atomic dimensions. Electrons with an orbital momentum different from zero contribute to the magnetic moment of an atom. In addition, a magnetic moment is arising from the spin of the electrons. Hence, if the expectation value of either the orbital momentum (L) or spin momentum (S) of an atom is different from zero, then the atom is magnetic. The term scheme of an atom for different combinations of L and S can be calculated by taking into account Coulomb interactions between electrons and the Pauli exclusion principle. The result shows that the ground state is obtained by filling of the atomic shells with electrons according to the **Hund's rules**

1. S has the maximum value, but must be compatible with the Pauli exclusion principle.
2. L has the maximum value, but must be compatible with the Pauli exclusion principle and rule 1.
3. $|L - S|$ has the minimal value for less than half filled shells or
   $L + S$ has the maximum value for more than half filled shells.

This leads in general to unfilled shells with a non-vanishing magnetic moment. Only those atoms with all L-subshells filled with $2(2L + 1)$ electrons are non-magnetic.
Upon incorporation into a solid the magnetic moment of most atoms is lost, and there are only very few magnetic solids. The main reason is the delocalization of the electrons due to the overlap of the atomic wave functions with those of neighboring atoms. The delocalization leads to a reduction of the kinetic energy and significantly contributes to the binding energy of crystals. This is particularly the case for the outer (s, p) valence electrons, but also for $3d$-electrons. The hybridization of electron orbitals of neighboring atoms causes a splitting of the energy levels, and only the energetically most favorable states are occupied. The relatively low energies of the multiplet splitting due to the second Hund's rule do not play a role anymore. Instead of maximizing L by successively filling the orbital states $m_l = -l, -l + 1, \ldots, l - 1, l$, all m_l-states are equally occupied. Hence, the orbital moment is reduced in the solid compared to the free atom. Another reason for the reduced orbital moment is the lower symmetry of the potential due to the neighboring atoms. The symmetry of the crystal potential is mainly determined by the nearest neighbor positions and is thus usually lower than the central potential of a single atom. As a consequence, the orbital momentum l is not a good quantum number anymore, and the eigenfunctions have to be indexed according to the symmetry group of the crystal. The combined effect of the reduced symmetry and the delocalization is a drastic reduction of complete quenching or the orbital moment.
These arguments hold for the 2^+-ions of the $3d$-series in ionic bound salts as shown in Fig. 1. We consider insulating salts as an intermediate situation between isolated atoms and atoms in a

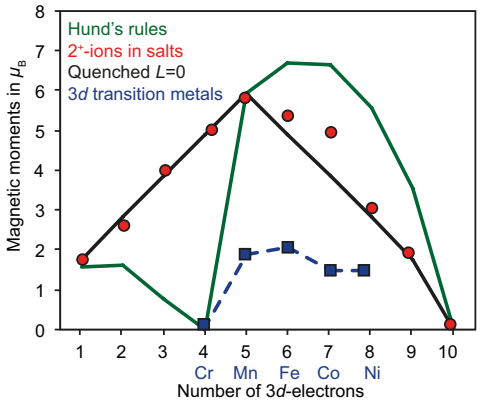


Fig. 1. *Comparison of measured effective magnetic moments (in μ_B) for 2^+-ions of the $3d$ series in salts (red symbols) with theoretical predictions for the pure spin moment, i.e. quenched L (black) and for total moment $L+S$ according to the Hund's rules, i.e. free atoms (green). The blue symbols indicate the magnetic moments per atom in the $3d$ transition metals Cr, Mn, Fe, Co, and Ni.*

metallic environment with strongly delocalized electrons, see below. The experimental values (red symbols) clearly deviate from the predictions of the Hund's rules (green curve), but are in rather good agreement with the assumption of complete quenching of the orbital moment $L = 0$ (black curve), while the intra-atomic exchange energy responsible for the first Hund's rule (S maximum) is still relevant. For $3d$ transition metals, however, the first Hund's rule breaks down, too. The magnetic moments per atom are strongly reduced compared to the free atoms, and only the elements in the middle of the $3d$-series (Cr, Mn, Fe, Co, and Ni) show magnetism (blue symbols), and the magnetic structure can be very complicated (*e.g.* non-commensurate antiferromagnetism in Cr). In these metals the picture of localized magnetic moments is not valid anymore. The $3d$-electrons are strongly delocalized and must be described in the band model (see Sec. 4). One speaks of *itinerant* electrons and correspondingly of *itinerant magnetism.*

The situation is different for the rare-earth elements, which show practically the same magnetic moments in the free atom, in ionic compounds with 3^+-ions, and in metals, see Fig. 2(a). The reason for this different behavior is the very strong localization of the $4f$-orbitals as exemplarily

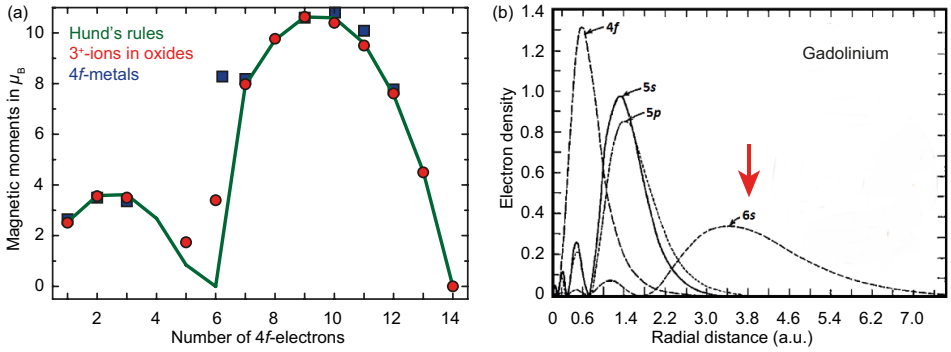


Fig. 2: *(a) Comparison of measured effective magnetic moments (in μ_B) for 3^+-ions of the $4f$ rare-earth series in R_2O_3 oxides (red symbols) and $4f$-metals with theoretical predictions according to the Hund's rules,* i.e. *for free atoms (green). (b) Radial electron density of the outer electrons in Gd. The red arrow indicates the nearest-neighbor distance in Gd metal.*

shown in Fig. 2(b) for Gd with the electron configuration [Xe]$4f^7 5d^1 6s^2$. The $4f$-orbitals are clearly closer to the core than the $5s$ and $5p$-orbitals, which belong to the [Xe] core. The overlap of the $4f$-orbitals with orbitals of the nearest neighbor atom [red arrow in Fig. 2(b)] is very weak. The $4f$-orbitals hardly contribute to the crystal bonding of Gd metal. Their atomic environment is weakly perturbed in the crystal, and the second Hund's rule remains a good approximation.

3 Interaction of magnetic moments

The overlap of electron orbitals of neighboring atoms in a solid gives rise to correlations of these electrons resulting in inter-atomic interaction. As a consequence, the total energy of the solid depends on the relative alignment of the magnetic moments of neighboring atoms. The interaction range and the type of interaction (*e.g.* ferromagnetic or antiferromagnetic) depends on the type of binding in the crystal (ionic, covalent, or metallic), *i.e.* the extent of the electron correlations. Different basic coupling mechanism are distinguished. However, for a given material a clear assignment to one of them is usually not possible, the transitions are gradual as are the binding types in crystals (Fig. 3).
Figure 3 schematically shows the situations for different types of exchange mechanisms in solids. Direct exchange requires direct overlap of (moment carrying) wave functions, indirect exchange is mediated by polarization of the electrons of the medium, and superexchange is mediated by intermediate ions. The main coupling types are briefly discussed in the following sections. In any case, exchange interaction is the consequence of solely Coulomb interaction in combination with the Pauli exclusion principle.

3.1 Dipole interaction

Classical dipole-dipole interaction

$$E_{\text{dip}}(\vec{m}_1, \vec{m}_2, \vec{R}) = \frac{\mu_0}{4\pi R^3}\left(\vec{m}_1 \cdot \vec{m}_2 - \frac{3}{R^2}(\vec{m}_1 \cdot \vec{R})(\vec{m}_2 \cdot \vec{R})\right) \tag{1}$$

is always present between two magnetic moments $\vec{m}_{1,2}$ at a distance $\vec{R}$ and depends on the relative orientation. The dipole-dipole interaction is way too weak to cause a cooperative alignment of the moments at typical Curie temperatures of ferromagnets. An estimation with $m_1 = m_2 = 1\,\mu_{\text{B}}$ and $R = 2\,\text{Å}$ yields $E_{\text{dip}} \approx 10^{-5}\,\text{eV} \approx 0.2\,\text{K}$. However, the dipole interaction will play an important role for the discussion of magnetic anisotropy (Sec. 7), in particular for the alignment of the magnetic moments in thin films.

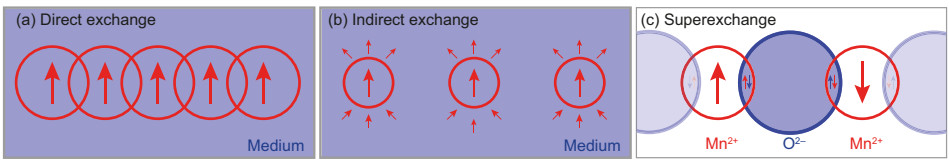


Fig. 3: *(a) Direct exchange due to overlap of wave functions, (b) indirect exchange mediated by a polarized medium, and (c) superexchange mediated by an intermediate ion (here O^{2-} in the antiferromagnetically ordered MnO solid).*

3.2 Direct exchange

For a basic understanding of exchange coupling it is instructive to consider the simplest 2-electron system, the H_2 molecule. The Hamilton operator of the electronic system is

$$\mathcal{H} = -\frac{\hbar^2}{2m}\partial^2_{\vec{r}_1} - \frac{e^2}{|\vec{r}_1 - \vec{R}_a|} - \frac{\hbar^2}{2m}\partial^2_{\vec{r}_2} - \frac{e^2}{|\vec{r}_2 - \vec{R}_b|} \quad (2)$$

$$+ \frac{e^2}{R} + \frac{e^2}{|\vec{r}_1 - \vec{r}_2|} - \frac{e^2}{|\vec{r}_1 - \vec{R}_b|} - \frac{e^2}{|\vec{r}_2 - \vec{R}_a|}, \quad (3)$$

where $\vec{R}_{a,b}$ and $\vec{r}_{1,2}$ denote the coordinates of the two H cores and the two electrons, respectively, and $R = |\vec{R}_a - \vec{R}_b|$ is the core-core distance. The first line (2) contains the terms of the two separated atoms a and b and the second line (3) the interaction $W(\vec{r}_1, \vec{r}_2)$ between the atoms. Note that the spin does not appear explicitly in the Hamilton operator. Hence, the wave functions of the two electrons can be written as products of the spin (χ) and spatial (Φ) parts. In addition, since electrons are Fermions the **Pauli exclusion principle** – the second ingredient apart from Coulomb interaction – must be fulfilled: Therefore the total wave function Ψ has to be antisymmetric under exchange of the two non-distinguishable electrons:

$$\Psi(\vec{r}_1, \vec{s}_1; \vec{r}_2, \vec{s}_2) = \phi(\vec{r}_1, \vec{r}_2)\chi(\vec{s}_1, \vec{s}_2) = -\Psi(\vec{r}_2, \vec{s}_2; \vec{r}_1, \vec{s}_1). \quad (4)$$

The antisymmetry required by the Pauli principle can be fulfilled in two ways

1. Singlet, spins antiparallel, $S = 0$:
   Antisymmetric spin wave function: $\chi_s(\vec{s}_1, \vec{s}_2) = -\chi_s(\vec{s}_2, \vec{s}_1)$
   Symmetric spatial wave function: $\phi_s(\vec{r}_1, \vec{r}_2) = \phi_s(\vec{r}_2, \vec{r}_1)$

2. Triplet, spins parallel, $S = 1$:
   Symmetric spin wave function: $\chi_t(\vec{s}_1, \vec{s}_2) = \chi_t(\vec{s}_2, \vec{s}_1)$
   Antisymmetric spatial wave function: $\phi_t(\vec{r}_1, \vec{r}_2) = -\phi_t(\vec{r}_2, \vec{r}_1)$

There is only one antisymmetric combination of spins of the two electrons (thus termed *singlet*), but three symmetric configurations (thus termed *triplet*). The singlet state has $S = 0$ and correspondingly $M_S = 0$, whereas the triplet state correspond to $S = 1$ with $M_S = +1, 0, -1$. The different symmetries of the associated charge densities schematically shown in Fig. 4 result in different eigenvalues E_s and E_t for singlet and triplet states. The charge density in the singlet state is enhanced between the nuclei compared to the simple superposition of the atomic charge densities, whereas the charge density for the triplet state is reduced between the nuclei. The charge accumulation in the singlet state leads to the binding of the molecule since the repulsive core-core interaction is screened and overcompensated. Hence, the singlet state is the ground state of the H_2 molecule. The energy difference $E_s - E_t$ is given by the exchange integral

$$E_s - E_t = 2J = 2\int d\vec{r}_1 d\vec{r}_2 \; \varphi_0(\vec{r}_1 - \vec{R}_a)\varphi_0(\vec{r}_2 - \vec{R}_a)\; W(\vec{r}_1, \vec{r}_2)\; \varphi_0(\vec{r}_1 - \vec{R}_b)\varphi_0(\vec{r}_2 - \vec{R}_b), \quad (5)$$

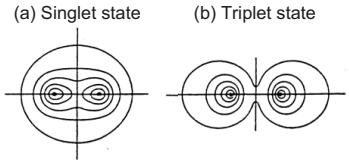


Fig. 4. *Schematic representation of the charge distributions of (a) the singlet and (b) the triplet state for the H_2 molecule.*

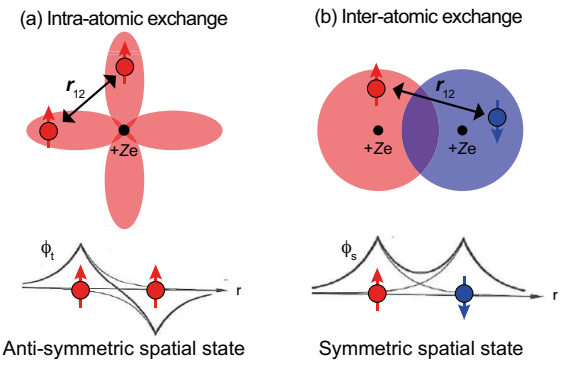


Fig. 5. *(a) Intra-atomic exchange leads to a triplet ground state with ferromagnetic exchange coupling $J > 0$. The antisymmetric spatial wave function minimizes the electron-electron and electron-core Coulomb interaction. (b) Inter-atomic exchange leads to the singlet ground state with $J < 0$. The symmetric spatial wave function forms bonding orbitals and thus (over)compensates the core-core repulsion.*

where $\varphi_0(\vec{r})$ is the $1s$ wave function of the hydrogen atom and $W(\vec{r}_1, \vec{r}_2)$ is the interaction between the atoms to be minimized [see Eq. (3)]. J is called **exchange coupling constant** of the two spins.

In general, beyond the H_2 molecule, the energy difference $E_s - E_t$ can be negative or positive favoring antiparallel spin alignment like in the H_2 molecule or parallel as required by the first Hund's rule for electrons within an atom, respectively. The difference between intra-atomic and inter-atomic exchange can be understood in the following way: The antisymmetric spatial wave functions of the electrons within the same atom (i) reduce their overlap thereby reducing the electron-electron Coulomb energy and (ii) reduce the mutual screening from the positively charged core thereby reducing the electron-core Coulomb energy [Fig. 5(a)]. The symmetric spatial wave function of electrons in neighboring atoms of a molecule or a solid increases the charge density between the atoms and forms bonds thereby reducing the Coulomb energy of the cores [Fig. 5(b)]. In the general case of a solid, electrons do not move in a central potential and the interaction term in Eq. (3) contains *many* electron-electron, core-core, and inter-atomic electron-core interactions. The relative weights of all these terms determine the sign of the exchange integral and the coupling constant J in Eq. (5).

3.3 Heisenberg spin Hamiltonian

If one is only interested in spin configurations and dynamics, an effective Hamiltonian can be constructed that solely acts in the spin space and yields the eigenvalues E_s and E_t. The projection operators $P_t = \frac{1}{2}\vec{S}^2$ and $P_s = 1 - P_t = 1 - \frac{1}{2}\vec{S}^2$ have the following properties:

$$P_t\chi_t^{(i)} = \chi_t^{(i)} \quad \text{and} \quad P_s\chi_t^{(i)} = 0 \quad (i = 1\ldots 3), \tag{6}$$

$$P_t\chi_s = 0 \quad \text{and} \quad P_s\chi_s = \chi_s, \tag{7}$$

where we use for the triplet ($S = 1$) states $\vec{S}^2\chi_t^{(i)} = S(S+1)\chi_t^{(i)} = 2\chi_t^{(i)}$ and for the singlet ($S = 0$) state $\vec{S}^2\chi_s = S(S+1)\chi_s = 0$. We can now write the effective Hamiltonian, which is diagonal in the spin space, as

$$\mathcal{H} = E_sP_s + E_tP_t. \tag{8}$$

Replacing $\vec{S}^2$ by

$$\vec{S}^2 = (\vec{s}_1 + \vec{s}_2)^2 = \vec{s}_1{}^2 + \vec{s}_2{}^2 + 2\vec{s}_1 \cdot \vec{s}_2 \tag{9}$$

and noting that $\vec{s}_i{}^2$ applied to spin functions with $s = \frac{1}{2}$ yields $s(s+1) = \frac{3}{4}$, we get

$$\mathcal{H} = \frac{1}{4}(E_s + 3E_t) - (E_s - E_t)\vec{s}_1 \cdot \vec{s}_2. \tag{10}$$

The first term in Eq. (10) is the average energy of all four possible spin combinations and the second part describes the dependence on the relative spin alignment. According to Eq. (5) the energy difference $E_s - E_t = 2J$ is given by the exchange integral. Therefore, the spin-dependent term of the effective Hamiltonian is usually written as

$$\mathcal{H}_{\text{spin}} = -2J\,\vec{s}_1 \cdot \vec{s}_2. \tag{11}$$

For positive (negative) J parallel (antiparallel) alignment of the spins is preferred, and correspondingly the coupling is called ferromagnetic for $J > 0$ and antiferromagnetic for $J < 0$.
The heuristic generalization of Eq. (11) derived for two localized electrons with spin $\frac{1}{2}$ to many spins $\vec{s}_n$ $(n = 1, \ldots, N)$, which can even be composed of several electron spins (*i.e.*, $s > \frac{1}{2}$), with empiric coupling constants J_{mn} between pairs of spins (n, m) at the positions $\vec{R}_n$ and $\vec{R}_m$ leads to the **Heisenberg operator for spin systems**

$$\mathcal{H}_{\text{spin}} = -2 \sum_{m \neq n} J(\vec{R}_n, \vec{R}_m)\,\vec{s}_n \cdot \vec{s}_m. \tag{12}$$

A stringent derivation of this operator and the calculation of the coupling constants from first principles has only partly been achieved so far. Nevertheless, this operator has manifold applications for magnetic systems with permanent magnetic moments. However for itinerant magnets such as $3d$ metals it is not suitable, see Sec. 4.

3.4 Indirect exchange

If the wave functions of the electrons that create the magnetic moment do not directly overlap, one speaks about **indirect exchange** [Fig. 3(b)]. In this case other, usually conduction electrons are polarized and mediate the coupling. One example are the $4f$-electrons of rare-earth metals. The strongly localized $4f$-electrons interact *via* polarizing the delocalized s, p-electrons [Fig. 2(b)] of the intermediate medium. Indirect exchange is typically weaker than direct exchange as manifested by the low Curie temperature of Gd of only 293 K.
Another example of indirect exchange are magnetic impurities in non-magnetic metals with delocalized s, p-electrons (acting as medium), where the indirect mechanism leads to the long-ranged **RKKY interaction** (named after the authors Rudermann, Kittel, Kasuga, and Yosida) with an oscillatory behavior as a function of the distance r between the impurities

$$J_{\text{RKKY}}(|\vec{r}|) \propto \frac{\cos(2k_F r)}{(2k_F r)^3}. \tag{13}$$

The oscillation period λ is given by the Fermi wave vector k_F of the medium, $\lambda = \pi/k_F$, and the strength decays with r^{-3}. The origin is a spin density wave (Friedel oscillations) created by each localized magnetic moment in the surrounding medium (delocalized electrons) [Fig. 6(a)]. The spin density wave causes indirect exchange at the relative position $\vec{r}$ of another localized moment. The RKKY mechanism is also the origin of the oscillatory **interlayer exchange coupling** observed in ferromagnet/non-magnet/ferromagnet multilayers with non-magnetic interlayers of a few nanometers thickness. The Hamiltonian of this interlayer coupling can be

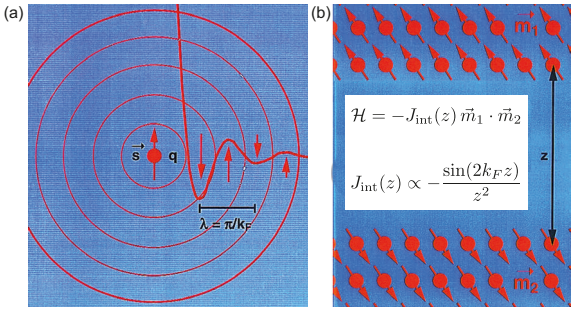


Fig. 6. *(a) Spin density wave of a localized moment (charge q and spin $\vec{s}$) in a medium of free or strongly delocalized electrons (blue). (b) Interlayer exchange coupling as a superposition of pairwise RKKY interactions.*

written in the Heisenberg form [Eq. (12)] proportional to the scalar product of the magnetizations $\vec{m}_{1,2}$ of the two ferromagnetic layers, see inset of Fig. 6(b). The coupling constant J_{int} is obtained by summing over all pair-wise RKKY interactions $J_{\text{RKKY}}(|\vec{r}|)$ [Eq. (13)] between localized moments in Fig. 6(b). The interlayer exchange coupling oscillates as a function of the interlayer thickness z and decays with z^{-2}. The oscillation period is completely determined by the Fermi wave vector k_F of the interlayer material and is of the order of only a few angstrom. In some cases, *e.g.*, Fe/Cr/Fe(100) trilayers, the sign of J_{RKKY} changes when the interlayer thickness is changed by one atomic layer causing the magnetizations $\vec{m}_{1,2}$ to switch from parallel to antiparallel alignment or *vice versa*. A more rigorous treatment of interlayer exchange coupling in terms of interference of a quantum-well state in the interlayer due to spin-dependent confinement confirms the simplified RKKY picture.

The discovery of antiferromagnetic interlayer exchange coupling by Peter Grünberg in 1986 [3] provided the novel possibility to control the relative alignment of spins separated by only a few nanometers with an external magnetic field. This triggered the first observations of the giant magnetoresistance (GMR) effect a few years later [4, 5] (Sec. 11), which then became the cornerstone of spintronics.

3.5 Superexchange

A special form of indirect exchange coupling is found in ionic and covalently bound insulators, *e.g.* oxides of transition metals and rare-earth elements (MnO, Fe_2O_3, Gd_2O_3). Figure 3(c) shows the example of MnO. The O^{2-}-ions are covalently bound to the two neighboring Mn^{2+}-ions. The p-shell of the O^{2-}-ion is completely filled, and all orbitals are occupied by a pair of spin-up and spin-down electrons. This is why the spins of the two electrons taking part in the bonding to the Mn^{2+}-ions have antiparallel alignment. The Mn^{2+}-ion has five $3d$-electrons, which according to the first Hund's rule all are ferromagnetically aligned. Therefore, the covalent bonding to a neighboring O^{2-}-ion is energetically favorable, if the $2p$-orbital of the O^{2-}-ion overlaps with antiparallel spin alignment [see red and blue arrows in the overlap region in Fig. 3(c)]. In total, there is a chain of three antiparallel spin alignments, which results in a net antiferromagnetic coupling between the moments of the two Mn^{2+}-ions. The range of superexchange interaction strengths lies between those of direct exchange and RKKY interaction.

3.6 Dzyaloshinskii-Moriya interaction

The exchange interactions discussed so far are isotropic, meaning that the coupling depends on the distance $|\vec{r}_{12}|$ between the spins but not on the orientation of $\vec{r}_{12}$ with respect to the crystalline environment. Anisotropic exchange takes the coupling to the crystal lattice into account but only occurs if the inversion symmetry is broken. This is the case in inversion asymmetric crystal structures or at interfaces and surfaces. In Figs. 7(b) and (c) the breaking of the inversion symmetry is sketched by considering a third atom (blue) in asymmetric positions with respect to $\vec{s}_1$ and $\vec{s}_2$. The electrostatic potential gradient due to the additional atom generates an electric field between $\vec{s}_1$ and $\vec{s}_2$ that gives rise to spin-orbit interaction (Sec. 5). The resulting **anisotropic exchange interaction** can be written as the so-called Dzyaloshinskii-Moriya interaction (DMI)

$$\mathcal{H}_{\mathrm{DM}} = \vec{D}_{12} \cdot \vec{s}_1 \times \vec{s}_2 \quad ; \quad \vec{D}_{12} \propto \xi\, \vec{a} \times \vec{r}_{12}. \tag{14}$$

$\vec{D}_{1,2}$ is the **Dzyaloshinskii-Moriya vector**, which is proportional to the spin-orbit coupling constant ξ and depends on the position of the third atom with respect to the spins $\vec{s}_1$ and $\vec{s}_2$ (Fig. 7). $\vec{D}_{1,2}$ points perpendicular to the triangle spanned by the three atoms. $\mathcal{H}_{\mathrm{DM}}$ favors configurations with perpendicular spins and induces a sense of rotation depending on the direction of $\vec{D}_{1,2}$, which according to Eq. (14) is related to the local atomic configuration given by the direction of $\vec{a}$. In the geometry of Fig. 7, if $\vec{D}_{1,2}$ points out of the drawing plane, the spins are canted in such a way that a clockwise (CW) rotation transforms $\vec{s}_1$ into $\vec{s}_2$ [Fig. 7(b)]. For reversed $\vec{a}$, $\vec{D}_{1,2}$ points into the drawing plane, and a counterclockwise (CCW) rotation connects $\vec{s}_1$ and $\vec{s}_2$. Note that pure DMI would lead to a perpendicular spin alignment, which minimizes Eq. (14). In Fig. 7 we implicitly assume dominant ferromagnetic direct exchange. The effect of the DMI in this case is to reduce the total magnetic moment and to induce an antiferromagnetic component (projections of $\vec{s}_1$ and $\vec{s}_2$ onto $\vec{r}_{12}$). If we assume antiferromagneitc direct exchange, DMI reduces the magnetization in the sublattices of the antiferromagnetic order and induces a weak ferromagnetic moment, which would point parallel (antiparallel) to $\vec{r}_{12}$ for the two directions of $\vec{D}_{1,2}$ in Figs. 7(b) and (c), respectively.

DMI was first postulated in the 1950s by Igor Dzyaloshinskii [6] and microscopically explained in terms of spin-orbit coupling in 1960 by Tori Moriya [7]. Later it was used to explain weak

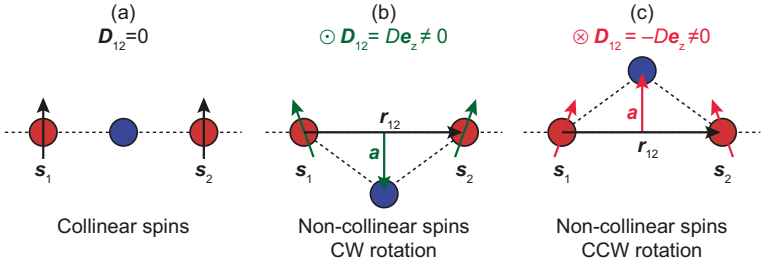


Fig. 7: *DMI is only effective when the inversion symmetry with respect to a plane containing $\vec{r}_{12}$ is broken. (a) $\vec{D}_{12} = 0$ for the symmetric case, the spins are collinear. (b,c) Breaking the symmetry leads to non-collinear spin alignment with the sense of spin rotation depending on the direction of the DMI vector $\vec{D}_{1,2}$. The blue atom in (b) and (c) represents a nearest-neighbor atom in a inversion asymmetric bulk material or a substrate atom in the case of a thin film structure. $\xi > 0$ is assumed in this figure.*

canting of the spins in antiferromagnetic materials, *e.g.* α-Fe_2O_3, $MnCO_3$, and $CoCO_3$. Only in recent years, DMI has again attracted increased attention, when it was found (i) to stabilize helical and Skyrmionic structures in ultra-thin films [8, 9] and in inversion asymmetric bulk materials (*e.g.* MnSi [10]), (ii) to give rise to a novel chiral magnetic domain wall structure in thin bilayer films [11], and (iii) to provide a mechanism for magnetism-induced electric polarization in a recently discovered class of multiferroics [12]. In the case of ultra-thin films, the interface to the substrate and the surface break the inversion symmetry, *i.e.* substrate atoms act as *third, blue atoms* in Fig. 7.

3.7 Itinerant exchange

If the electrons contributing to the magnetization are delocalized, then all coupling considerations have to be done in the band picture of the electronic structure. Also in this case, correlations of the electrons due to their spin alignment can lead to energetically favored spin alignments. Since this kind of *itinerant* exchange must be applied to describe the most important magnetic metals like Fe, Co, Ni the middle of the $3d$-series and their alloys, it will be discussed in more detail in Sec. 4.

4 Band magnetism (Itinerant magnetism)

As discussed in Sec. 2 most atoms are magnetic, but in solids magnetism is rather the exception. The reason is that parallel alignment of spins on the one hand gains exchange energy, but on the other hand can lead to a strong increase of kinetic energy. This is a direct consequence of the delocalization of the valence electrons when going from the atom to the solid and can be discussed in the model of free electrons. In the non-magnetic case, all states $\vec{k}$ inside the Fermi sphere with radius k_F are doubly occupied with two electrons of opposite spin. $k_F = (3\pi^2 n)^{1/3}$ is given by the electron density $n = N/V$. The sum of the kinetic energies of all electrons is $E_{kin} = N\frac{3}{5}\frac{\hbar^2}{2m}k_F^2$. In the completely ferromagnetic case, where all states are occupied with only one spin-up electron, the volume of the Fermi sphere doubles, k_F increases by $2^{1/3}$ and E_{kin} by $2^{2/3} = 1.587$. This strong increase cannot be compensated by exchange energy. Thus, completely delocalized electrons do not tend to magnetism. In solids the degree of localization of the valence electrons determines whether the ground state is magnetic or not.
In order to develop a simple model for band ferromagnetism we consider **the spin-resolved density of states (DOS)** of the free-electron system $D(E) = D^{\uparrow}(E) + D^{\downarrow}(E)$ and assume that

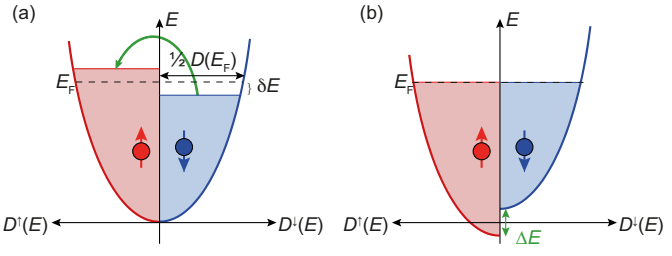


Fig. 8. *(a) Redistribution of spin-down electrons into unoccupied spin-up states. The redistribution increases the kinetic energy of the system. (b) Exchange-split DOS taking into account the increase in kinetic energy and the gain in exchange energy.*

some spin-down electrons are redistributed into unoccupied spin-up states, see Fig. 8(a). This leads to an increase of the kinetic energy of each redistributed electron by δE. The number of spin-up (spin-down) electrons is increased (decreased) by

$$\delta N = \frac{1}{2} D(E_{\mathrm{F}}) \delta E, \tag{15}$$

and the kinetic energy density (per volume V) in the system is increased by

$$\Delta E_{\mathrm{kin}} = \frac{\delta N}{V} \delta E = \frac{1}{2V} D(E_{\mathrm{F}}) (\delta E)^2. \tag{16}$$

With the total number of electrons per volume for the two spin orientations

$$n_{\uparrow,\downarrow} = \frac{N}{2V} \pm \frac{1}{2V} D(E_{\mathrm{F}}) \delta E \tag{17}$$

the unbalanced system has the magnetization

$$M = -\mu_{\mathrm{B}} (n_\uparrow - n_\downarrow) = -\mu_{\mathrm{B}} \frac{D(E_{\mathrm{F}})}{V} \delta E. \tag{18}$$

The minus sign reflects that the magnetic moment of an electron is opposite to its spin.
What are the conditions under which the formation of this magnetization and the related increase in kinetic energy can be compensated by a decrease of the Coulomb (or exchange) energy such that the total energy of the system is lowered? Each pair of electrons with parallel spin gains exchange energy $2I$ (I for each electron), because the antisymmetric spatial wave function $\phi_t(\vec{r}_1, \vec{r}_2)$ (compare Sec. 3.2) avoids that the two electrons are occasionally in the same orbital: $\phi_t(\vec{r}_1, \vec{r}_2) \to 0$ for $\vec{r}_1 \to \vec{r}_2$. This is not the case for the symmetric spatial wave function $\phi_s(\vec{r}_1, \vec{r}_2)$ of pairs of electrons with antiparallel spins. Hence, I can be identified with the exchange integral of an electron [similarly as in Eq. (5)]. In the present context I is called the **Stoner parameter**. The correlation effects among electrons of the same spin direction renormalize the single electron levels. Since the exchange interaction favors the parallel spin alignment, the majority DOS $D^\uparrow(E)$ is shifted down by $-\frac{1}{2}\Delta E$ and the minority DOS $D^\downarrow(E)$ up by $\frac{1}{2}\Delta E$:

$$D^{\uparrow(\downarrow)} = D(E \mp \frac{1}{2}\Delta E) \quad \text{with} \quad \Delta E = I(N_\uparrow - N_\downarrow). \tag{19}$$

This is called **exchange splitting** of the DOS of a ferromagnet, see Fig. 8(b). The number of electron pairs with parallel spins in the up(down) direction is (without correcting for the $N \ll N^2$ self-pairs)

$$\frac{1}{2} N^2_{\uparrow\uparrow(\downarrow\downarrow)} \approx \frac{1}{2} \left(\frac{N}{2} + (-) \delta N \right)^2 \tag{20}$$

yielding a gain in Coulomb energy density

$$\Delta E_{\mathrm{C}} \approx -\frac{2I}{V} \left(\frac{N^2}{4} + (\delta N)^2 \right). \tag{21}$$

The first term is the energy gain due to correlations for equal occupation of the two spin directions and the second due to the redistribution of electrons. Rewriting the second term using Eq. (17) the change of the Coulomb energy density due to the redistribution only becomes

$$\Delta E_{\mathrm{C}} = -\frac{I}{2V} D(E_{\mathrm{F}})^2 (\delta E)^2, \tag{22}$$

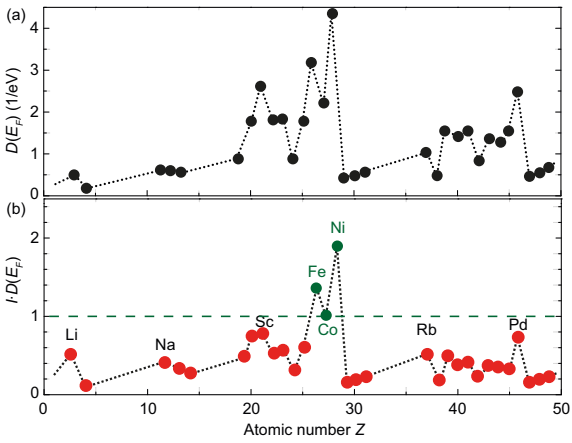


Fig. 9. *(a) $D(E_F)$ per atom and (b) the product $I \cdot D(E_F)$ as a function of the atomic number Z (data from Ref. [13]). Only Fe, Co, and Ni fulfill the Stoner criterion (above the dashed green line).*

which is proportional to M^2 [Eq. (18)]. With Eqs. (22) and (16) the total change of the energy density due to the redistribution becomes

$$\Delta E = \Delta E_{\mathrm{kin}} + \Delta E_{\mathrm{C}} = \frac{1}{2V} D(E_{\mathrm{F}})(\delta E)^2 \left[1 - I \cdot D(E_{\mathrm{F}})\right] . \tag{23}$$

The term in the bracket implies that a $I \cdot D(E_{\mathrm{F}}) > 1$ is a sufficient condition for band ferromagnetism called **Stoner criterion**. Ferromagnetism is favored for large exchange integral I and more importantly for large DOS at the Fermi level. The DOS of solids is in general strongly structured. In a simple approximation one can assume that the DOS of a band D scales inversely with the width W of the band, since the integral over the DOS of the band ($\approx W \cdot D$) is the total number of states in the band, *i.e.* a constant. The stronger the localization, the lower W and the higher D. In the atomic limit, W approaches zero, the Stoner criterion is always fulfilled, and the magnetic moment is maximum in accordance with the first Hund's rule.

4.1 Ferromagnetic 3d-metals: Fe, Co, Ni

Figure 9 shows calculated $D(E_{\mathrm{F}})$ and products $I \cdot D(E_{\mathrm{F}})$ relevant for the Stoner criterion for a variety of metals with atomic number Z. Obviously, the variation of the product $I \cdot D(E_{\mathrm{F}})$ is

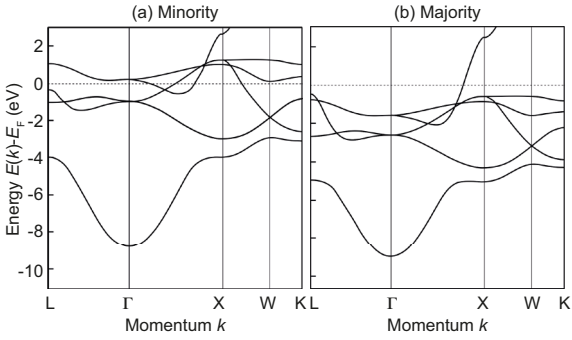


Fig. 10. *Bandstructures for (a) minority and (b) majority electrons for Co obtained from spin density functional calculations [14].*

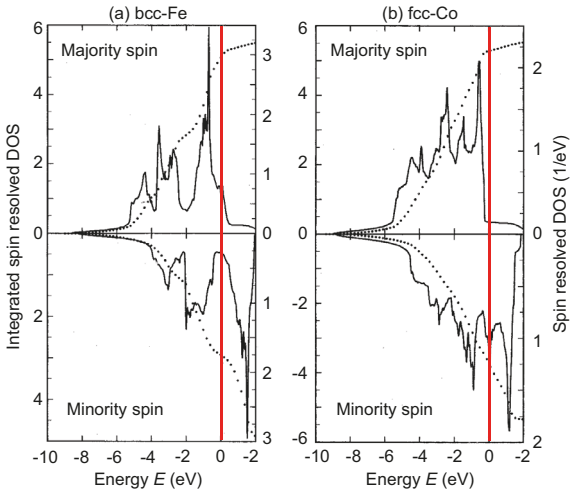


Fig. 11. *Spin-resolved DOS for (a) bcc-Fe and (b) fcc-Co obtained from spin density functional calculations [15]. The DOS of the majority and minority electrons (right scale) is plotted upward and downward, respectively. The states below the Fermi level $E = 0$ (red lines) are occupied. The dotted curves represent energy integrated density of states (left scale).*

mainly due to $D(E_\mathrm{F})$ and much less due to I. Only Fe, Co, and Ni fulfill the Stoner criterion and are indeed ferromagnetic. The figure indicates that Pd is at the verge to ferromagnetism, which is in agreement with the fact that a magnetic moment can be induced in Pd, if it is in proximity to a ferromagnetic metal.

Figure 10 shows exemplarily the bandstructures of minority and majority electrons in Co and Fig. 11 the spin-resolved DOS for Fe and Co that are dominated by $3d$-states. The majority and minority DOS are apart from the relative shift by the exchange splitting ΔE approximately the same, showing the applicability of the Stoner model for these real metals. The spin-resolved DOS of fcc-Ni looks very similar to that of fcc-Co except for a smaller exchange splitting ΔE. For Co and Ni the majority states are fully occupied, whereas 1.7 d-electrons for Co and 0.6 d-electrons for Ni are missing in the minority states leading to a magnetic moment per atom of 1.7 and 0.6 μ_B for Co and Ni, respectively. The magnetic moment per atom for Fe is 2.2 μ_B, however, the majority d-states are not fully occupied. Therefore, Fe is called a **weak magnet** in contrast to the **strong magnets** Co and Ni.

4.2 Antiferromagnetic $3d$-metals: *e.g.* Mn and Cr

Antiferromagnetic metals like γ-Mn with alternating positive and negative atomic moments along the cubic (001) axis or bcc-Cr exhibiting an incommensurable spin density wave can also be treated in a manner similar to the Stones model. The qualitative result is that antiferromagnetism is favored, if the Fermi level is in the middle of a band and $D(E_\mathrm{F})$ is small. The minority and majority DOS are equal, but some band crossings close to the Fermi level in the non-magnetic state are removed by hybridization leading to the opening of gaps in the antiferromagnetic case. The corresponding occupied bands are shifted to lower energy near the avoided crossings, which in total causes a gain in kinetic energy and thus the stabilization of the antiferromagnetic state.

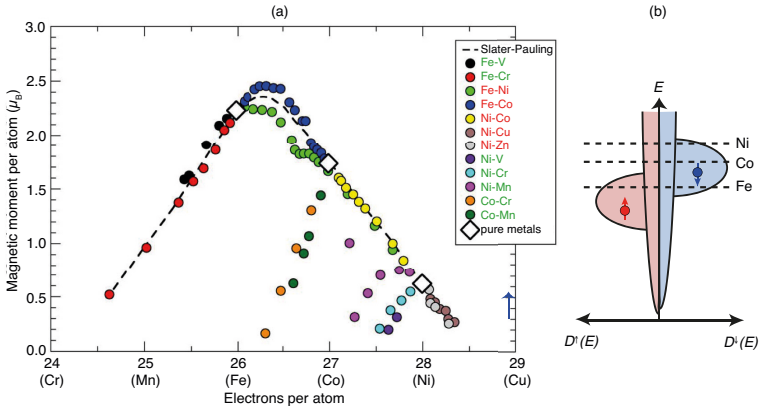


Fig. 12: *(a) Slater-Pauling plot for alloys of 3dmetals showing the mean magnetization per atom as a function of the mean number of electrons per atom. Adapted from Ref. [16]. (b) The rigid band model assumes that a variation of the number of electrons per atom can be described by moving the Fermi level, while maintaining a rigid DOS. s and p-states are schematically shown as wide parabolic bands, and the d-band as exchange-split narrow bands.*

4.3 Ferromagnetic 3d-alloys

The ferromagnetic 3d-metals Fe, Co, and Ni can be alloyed with other 3d transition metal elements yielding a rich variety of magnetic properties. In particular the magnetization of the alloys can be tailored. The magnetization increases (decreases) if the local atomic moment of the atoms added aupon alloying is parallel (antiparallel) to the moments of the host material. A compilation of the mean magnetic moments per atom of binary 3d-alloys as a function of the mean number of electrons per atom yield so-called **Slater-Pauling curve** in Fig. 12(a). The curve consists of two main branches with opposite slopes. Fe alloys lie on the left branch, whereas Co and Ni alloys form the right branch with negative slope and the side branches with positive slope. In Co and Ni hosts as strong ferromagnets valence electrons added upon alloying mainly occupy minority states and, thus, reduce the mean magnetic moment yielding the slope -1. In weak magnets like Fe the added electron can contribute to the majority or minority DOS. Since the DOS at the Fermi level is larger for majority than minority electrons [Fig. 11(a)], the magnetization increases with a positive slope. This behavior can easily be understood by assuming for simplicity a rigid DOS and a shifting Fermi level to account for the varying number of electrons per atom [Fig. 12(b)].

5 Spin-orbit coupling

In the previous sections the focus was on spin-spin interaction, while spin-orbit coupling (SOC) was largely neglected. This approximation allowed us to formulate the Heisenberg spin Hamiltonian [Eq. (12)] that only acts in the spin space. Due to the separation of the Hilbert space into orthogonal subspaces spanned by spatial and spin coordinates, respectively, the total wave functions could be factorized into spatial and spin parts, *e.g.* in Eq. (4). As a consequence, the spin s and magnetic quantum number m are good quantum numbers characterizing the eigenstates

of the total Hamiltonian. The inclusion of SOC changes this situation, since it mixes spin and spatial variables as we will see below.
SOC describes the interaction of the electron spin momentum with electrical fields that can be of internal (Coulomb interaction with the cores and other electrons) or external origin. In a classical picture, an electron moving with velocity $\vec{v}$ in an electric field $\vec{E}$ experiences in its own reference frame the Lorentz-transformed field that contains a magnetic field component $\vec{B} \approx \frac{1}{c^2}\vec{E} \times \vec{v}$ (for small v compared to the velocity of light c) that couples to the electron's magnetic moment. In quantum mechanics it is not obvious how to define a reference frame of a moving atom, but SOC directly follows from the Dirac equation as a relativistic effect (as reflected in the classical picture by the Lorentz transformation). When the Dirac equation is simplified to the Schrödinger equation with relativistic effect corrections added to the Hamiltonian, SOC is described by

$$\mathcal{H}_{\text{SOC}} = \frac{e\hbar}{4m_{\text{e}}^2c^2}\vec{E} \cdot (\vec{p} \times \vec{\boldsymbol{\sigma}}), \tag{24}$$

where m_{e} is the electron mass, $\vec{p}$ the linear momentum operator, and $\vec{\boldsymbol{\sigma}} = (\sigma_x, \sigma_y, \sigma_z)$ is a vector with the components being the Pauli matrices. The electric field can in general be written as $\vec{E} = -\vec{\nabla}V(\vec{r})$, where $V(\vec{r})$ is the electrostatic potential.
For a central potential, $V(\vec{r}) = V(r)$, the SOC Hamiltonian can be written as

$$\mathcal{H}_{\text{SOC}} = \frac{e\hbar}{4m_{\text{e}}^2c^2}\frac{1}{r}\frac{dV(r)}{dr}\vec{L} \cdot \vec{\boldsymbol{\sigma}} = \xi(r)\vec{L} \cdot \vec{\boldsymbol{S}} \quad \text{with} \quad \xi(r) = \frac{e\hbar}{2m_{\text{e}}^2c^2}\frac{1}{r}\frac{dV(r)}{dr}, \tag{25}$$

where $\vec{\boldsymbol{S}} = \vec{\boldsymbol{\sigma}}/2$ is the spin operator, and $\xi(r)$ is called the **spin-orbit couplng constant**. In this form the coupling of the spin with the angular momentum $\vec{L} = \vec{r} \times \vec{p}$ is directly evident. In atoms as well as in solids, a central potential is a meaningful first approximation because the Coulomb potential of the nucleus dominates, in particular in the vicinity of the nucleus that contributes most to Eq. (25), $\xi(r) \propto (1/r)dV/dr \sim -Z|e|/r^3 \propto -Z^4$, since the expectation value $\langle 1/r^3 \rangle \propto Z^3$ with Z being the atomic number. This strong dependence on Z is the reason why heavy atoms in general show stronger SOC than light atoms. The appearance of $\vec{L}$ in Eq. (25) implies that s-electrons with $L = 0$ are not expected to show SOC. SOC is especially strong for p-orbitals, which are closer to the nucleus than d or f-orbitals such that the r^{-3} dependence of $\xi(r)$ prevails the larger orbital momentum of the d or f-orbitals.
The product $\vec{L} \cdot \vec{\boldsymbol{\sigma}}$ in Eq. (25) can rewritten using the raising and lowering operator $L_\pm = L_x \pm iL_y$ and $\boldsymbol{\sigma}_\pm = \boldsymbol{\sigma}_x \pm i\boldsymbol{\sigma}_y$ for the z-component of the angular momentum and spin, respectively,

$$\vec{L} \cdot \vec{\boldsymbol{\sigma}} = L_z \cdot \boldsymbol{\sigma}_z + \frac{1}{2}(L_+\boldsymbol{\sigma}_- + L_-\boldsymbol{\sigma}_+). \tag{26}$$

The first term is spin conserving, while the second describes spin flipping. For instance, the action of the operator $\vec{L} \cdot \vec{\boldsymbol{\sigma}}$ on the wave function $|l, m > | \uparrow >$ adds an usually small component $|l, m+1 > | \downarrow >$ with opposite spin to the initial state due to the term $L_+\boldsymbol{\sigma}_-$. Thus, the action of SOC on a state with pure spin and pure angular momentum leads to a beating between higher and lower m and up and down spin: m and s are not constant anymore. Only this admixture of components with opposite spin character in the eigenstates allows for spin-flip processes. Since lowering (raising) the spin is accompanied by raising (lowering) of l, the total z-component m_j of the total angular momentum j is constant. The beating does neither change the angular momentum l. In other words, the operators L^2 and $\vec{J} = \vec{L} + \frac{1}{2}\vec{\boldsymbol{\sigma}}$ commute with $\mathcal{H}_{\text{SOC}}$, but L_z and $\boldsymbol{\sigma}_z$ don't.

Beyond the approximation of a central potential $V(r)$, the electric field $\vec{E}$ entering the SOC in Eq. (24) is determined by the details of the electrostatic potential landscape $V(\vec{r})$ that can give rise to a variety of SOC-based effects. Depending on the origin of the electric field $\vec{E}$ one can roughly distinguish

- **Symmetry-independent SOC effects** that exist in all types of crystals. The electric field stems from the intra-atomic Coulomb potential. Examples are magnetocrystalline anisotropy (Sec. 7), anisotropic magnetoresistance (Sec. 10), spin Hall effect, spin relaxation, and spin dependence of the Mott scattering cross section,

- **Symmetry-dependent SOC effects** that only occur in systems with broken inversion symmetry. The inversion symmetry may arise from the crystal structure in the bulk of a material (*e.g.* zinc-blende or B20 structure) and is called **Dresselhaus SOC** or from the symmetry breaking at surface and interfaces (*e.g.* thin films, heterostructures) and is referred to as **Bychkov-Rashba SOC**. Examples are Dzyaloshinskii-Moriya interaction (Sec. 3.6) in the bulk or in thin films, the Rashba effect in two-dimensional electron gas systems in asymmetric quantum well heterostructures, and the Rashba splitting of surface states of metals and semimetals (see below).

As a further example contrasting the intra-atomic central potentials discussed above, we consider a two-dimensional electron gas in the x-y-plane and a spatially constant electric field $\vec{E} = E_z\hat{z}$. This situation is approximately realized for surface states at metal surfaces, where the symmetry breaking gives rise to a potential gradient in the direction normal to the surface (z-direction), or in a two-dimensional electron gas in a semiconductor heterostructure. The Hamiltonian consists of the kinetic energy and the SOC term calculated according to Eq. (24) for $\vec{E} = E_z\hat{z}$

$$\mathcal{H} = \frac{\hbar^2}{2m_\mathrm{e}}k_\parallel^2 + \alpha_\mathrm{BR}\vec{\boldsymbol{\sigma}} \cdot (\vec{k_\parallel} \times \hat{z}), \tag{27}$$

where α_BR is the strength of the Bychkov-Rashba SOC. This SOC term can be interpreted as the interaction of the spin with an effective in-plane magnetic field, which is always perpendicular to the propagation direction $\vec{k_\parallel}$. For $\vec{k_\parallel} = (k_x, k_y, 0) = k_\parallel(\cos\varphi, \sin\varphi, 0)$ the eigenfunctions can be written as a product of plane waves times a two-component spinor

$$\psi_{\pm\vec{k}_\parallel}(\vec{r}_\parallel) = \frac{e^{i\vec{k}_\parallel\cdot\vec{r}_\parallel}}{2\pi}\frac{1}{\sqrt{2}}\begin{pmatrix} ie^{-i\varphi/2} \\ \pm e^{i\varphi/2} \end{pmatrix} \tag{28}$$

with eigenenergies

$$\varepsilon_\pm(\mathbf{k}_\parallel) = \frac{\mathbf{k}_\parallel^2}{2m_\mathrm{e}} \pm \alpha_\mathrm{BR}|\mathbf{k}_\parallel| = \frac{\hbar^2}{2m_\mathrm{e}}(k_\parallel \pm k_\mathrm{SO})^2 - \Delta_\mathrm{SO}\,, \tag{29}$$

where $\pm$ denotes spin-up and spin-down states with respect to the local (in $\vec{k}_\parallel$ space) spin orientation axis. The two-fold degenerate energy paraboloid of the free-electron model is now spin-split. The splitting $\varepsilon_+(\vec{k_\parallel}) - \varepsilon_-(\vec{k_\parallel}) = 2\alpha_\mathrm{BR}k_\parallel$ is linear in $k_\parallel$. The spin-split parabolas are shifted on the $k_\parallel$-axis by $k_\mathrm{SO} = m_\mathrm{e}\alpha_\mathrm{BR}/\hbar^2$ in opposite directions for spin-up and spin-down states, and the energy is overall lowered by $\Delta_\mathrm{SO} = m_\mathrm{e}\alpha_\mathrm{BR}^2/(2\hbar^2)$ as shown in Fig. 13. The so-called **Rashba splitting** of Au(111) surface states has been observed first by spin-averaged angle-resolved photoemission spectroscopy (ARPES) [17] and later by spin-resolved ARPES [18].

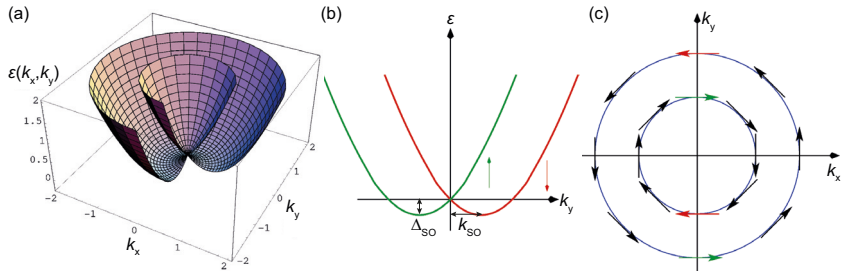


Fig. 13: *(a,b) Spin-split parabolic energy dispersion of a two-dimensional electron gas due to SOC in an inversion asymmetric environment. (c) Fermi surface with the arrows indicating the local quantization axis corresponding to the spin pattern.*

The spin pattern at the Fermi surface shown in Fig. 13(c) has interesting consequences on transport properties. Electrons with Fermi energy propagating for instance in the direction $\vec{k}_y$ have two different k-vectors for spin-up and spin-down components of the spinor [red and green in Figs. 13(b) and (c)] resulting in spin precession along the propagation direction. This feature of the Rashba splitting is exploited in the spin transistor proposed by Datta and Das [19], where a gate electrode in the heterostructure below the transport channel is used to modulate the electric field acting on the two-dimensional electron gas and, thus, the Rashba splitting and the spin precession.

6 Collective magnetism

The interactions between magnetic moments discussed in Sects. 3 and 4 give rise to collective magnetism and different magnetic ground states in a solid. In the ferromagnetic ground state all magnetic moments are aligned parallel. The antiferromagnetic state consists of two ferromagnetically ordered sublattices with opposite spin directions resulting in antiparallel alignment of adjacent moments. The total spontaneous magnetization vanishes. In a ferrimagnet the magnitude of the antiparallel aligned moments of the two sublattices are different resulting in a net total magnetization. Spin glasses are characterized by magnetic moments that are frozen out with random orientation. An example are magnetic impurities in a non-magnetic metallic host. The superposition of oscillatory and long-range pair-wise RKKY interactions [Eq. (13)] cause the random orientation of the impurities' moments. Finally, there are helical and spiral or even chiral magnetization arrangements that are stabilized by suitable combinations of nearest and next-nearest Heisenberg-type interactions [Eq. (11)], *e.g.* in rare-earth metals with hcp structure, or by Dzyaloshinskii-Moriya interaction [Eq. (14)], *e.g.* for a single atomic layer of Mn on a W(110) substrate [8]. In the following we will focus on ferro-, ferri-, and antiferromagnetic order that can be treated in the mean-field approximation to describe the temperature dependence of the magnetization and the magnetic susceptibility, *i.e.* the response to an externally applied magnetic field.

6.1 Ferromagnetic order

Ferromagnetism is characterized by a spontaneous magnetization even in the absence of an external magnetic field, which however only occurs below a material-dependent temperature T_C called **Curie temperature**. Above T_C thermal fluctuations destroy the magnetic order. Starting from the Heisenberg model [Eq. (11)] with constant nearest-neighbor interaction $J_\mathrm{A} > 0$, the interaction of a given atomic site $\vec{J}_i$ with its z neighbors is

$$E_i = -2J_\mathrm{A} \sum_{j=1}^{z} \vec{J}_i \cdot \vec{J}_j. \tag{30}$$

We now consider mean values and replace the $\vec{J}_j$ by their time-averaged mean value $\langle \vec{J}_j \rangle$ and obtain the mean exchange energy

$$E_i = -2zJ_\mathrm{A} \langle \vec{J}_j \rangle \cdot \vec{J}_i. \tag{31}$$

With $\vec{M} = -ng_\mathrm{J}\mu_\mathrm{B}\langle \vec{J}_j \rangle$ (atomic density n, Landé factor g_J, Bohr magnetron μ_B)

$$E_i = -(-g_\mathrm{J}\mu_\mathrm{B}\vec{J}_i) \cdot \frac{2zJ_\mathrm{A}}{ng_\mathrm{J}^2\mu_\mathrm{B}^2}\vec{M} = -\vec{\mu} \cdot \vec{B}_\mathrm{A}. \tag{32}$$

Hence, we have *formally* described the exchange energy as the product of the magnetic moment $\vec{\mu}$ and an *effective* magnetic field $\vec{B}_\mathrm{A}$ called **exchange field** or **molecular field**, which is proportional to the magnetization $\vec{B}_\mathrm{A} = \mu_0\lambda\vec{M}$. λ is called **molecular field constant** and is positive for ferromagnets. This procedure is the **mean-field approximation**. The molecular field is a fictitious mean field that creates in a solid the same order as the exchange interaction. In the case of an externally applied magnetic field, the effective field $\vec{B}_\mathrm{eff} = \vec{B}_\mathrm{A} + \vec{B}_\mathrm{ext} = \mu_0\lambda\vec{M} + \vec{B}_\mathrm{ext}$ must be considered. At $T = 0$ all $\vec{\mu}$ are aligned parallel to $\vec{B}_\mathrm{eff}$. All magnetic quantum numbers are $m_J = -J$ minimizing the energy $-\mu_z B_\mathrm{eff}$, and the magnetization equals the saturation magnetization M_S.
At finite temperatures thermal excitations populate levels with $m_J > -J$ with a higher energy $-\mu_z B_\mathrm{eff}$. The temperature dependence of the magnetization $\vec{M}(T)$ is obtained by averaging all $m_j = -J \cdots + J$ possible alignments of $\vec{\mu}$ with respect to $\vec{B}_\mathrm{eff}$ weighted by the corresponding occupation probability given by the Boltzmann factor $\exp(m_J g_\mathrm{J}\mu_\mathrm{B}B_\mathrm{eff}/k_\mathrm{B}T)$. The result is

$$\frac{M(T)}{M_\mathrm{S}} = B_J\left(\frac{g_\mathrm{J}\mu_\mathrm{B}J(\mu_0\lambda M(T) + B_\mathrm{ext})}{k_\mathrm{B}T}\right) = B_J(y), \tag{33}$$

with

$$y = \frac{g_\mathrm{J}\mu_\mathrm{B}J(\mu_0\lambda M(T) + B_\mathrm{ext})}{k_\mathrm{B}T}. \tag{34}$$

$B_J(y)$ is the **Brillouin function** and k_B the Boltzmann constant. Equation (33) is an implicit function of $M(T)$. In order to find non-trivial solutions ($M \neq 0$) we solve Eq. (34) for M,

$$\frac{M(T)}{M_\mathrm{S}} = \frac{k_\mathrm{B}T}{\mu_0 g_\mathrm{J}\mu_\mathrm{B}J\lambda M_\mathrm{S}}y - \frac{B_\mathrm{ext}}{\mu_0\lambda M_\mathrm{S}}, \tag{35}$$

and plot the two expressions for $M(T)/M_\mathrm{S}$ [Eqs. (33) and (35)] as a function of y in Fig. 14. The black curves show saturation for large $|y|$. *i.e.* large $\vec{B}_\mathrm{eff}$ and small T, and $M = 0$ for $y = 0$,

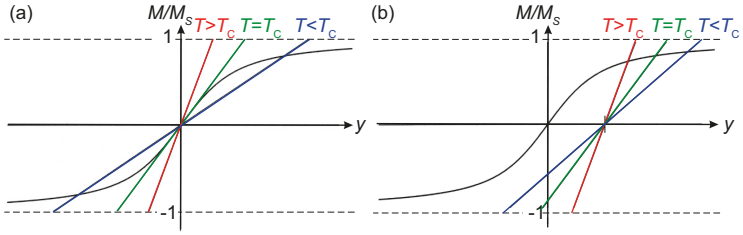


Fig. 14: *Graphical determination of $M(T)$ from Eqs. (33) and (35) for (a) $B_{ext} = 0$ and (b) $B_{ext} \neq 0$.*

i.e. small $\vec{B}_{\text{eff}}$ and large T, where thermal fluctuations destroy the magnetic order. A solution for M exists, if the two curves intersect. For $\vec{B}_{\text{ext}} = 0$ in Fig. 14(a) this is only the case, if the temperature is lower than a critical value T_{C} that can be identified as the Curie temperature

$$T_{\text{C}} = \frac{\mu_0 g_{\text{J}}(J+1)\mu_{\text{B}}\lambda M_{\text{S}}}{3k_{\text{B}}} = \lambda C, \tag{36}$$

where C is the material-dependent **Curie constant**. Equation (36) confirms the intuitive expectation that materials with strong exchange interaction exhibits a high Curie temperature and allows to estimate the strength of the molecular field $B_{\text{A}} = \lambda M_{\text{S}}$. For $J = 1/2$, $g_{\text{J}} = 2$, and $T_{\text{C}} = 1000\,\text{K}$ we obtain $B_{\text{A}} \approx 1500\,\text{T}$. This extremely high value illustrates the strength of the exchange interaction. Equation (33) can now be rewritten (for $\vec{B}_{\text{ext}} = 0$) as

$$\frac{M}{M_{\text{S}}} = B_J\left(\frac{\mu_0 g_{\text{J}}\mu_{\text{B}} J\lambda M(T)}{k_{\text{B}}T}\right) = B_J\left(\frac{3J}{J+1}\cdot\frac{M}{M_{\text{S}}}\cdot\frac{T_{\text{C}}}{T}\right) \tag{37}$$

and is shown in Fig. 15(a) for different values of J and zero external field. The magnetization drops continuously with temperature and vanishes at T_{C}.

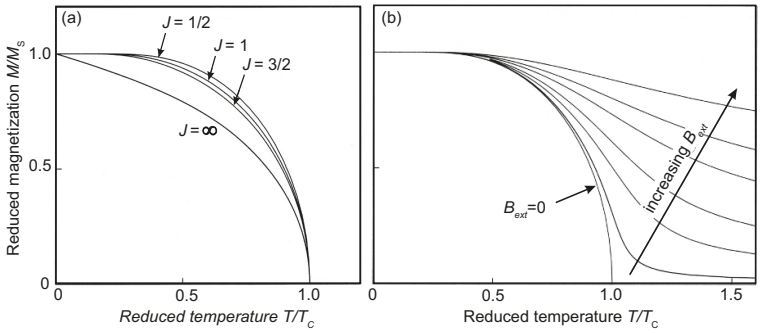


Fig. 15: *Relative magnetization as a function of the reduced temperature T/T_C (a) for different values of J and (b) for increasing external field B_{ext} for fixed $J = 1/2$. In the classical limit $J \to \infty$ the Brillouin function B_J in (a) approaches the Langevin function.*

For the discussion of the **magnetic susceptibility** $\chi = \mu_0 \frac{\delta M}{\delta B}$ we consider a small external field B_{ext} applied at $T > T_{\text{C}}$ that induces a small magnetization M. For small y, $B_J(y) \approx \frac{J+1}{3J} y$ and Eq. (33) becomes

$$\frac{M}{M_{\text{S}}} \approx \frac{J+1}{3J} \cdot \frac{g_{\text{J}} \mu_{\text{B}} J(\mu_0 \lambda M + B_{\text{ext}})}{k_{\text{B}} T} = \frac{T_{\text{C}}}{\mu_0 \lambda M_{\text{S}}} \frac{\mu_0 \lambda M + B_{\text{ext}}}{T}. \tag{38}$$

Solving for $\frac{M}{M_{\text{S}}}$ we get

$$\frac{M}{M_{\text{S}}} = \frac{T_{\text{C}}}{T} \cdot \frac{B_{\text{ext}}}{\lambda M_{\text{S}}} \cdot \left(1 - \frac{T_{\text{C}}}{T}\right)^{-1} \quad \text{and hence} \quad \chi = \frac{C}{T - T_{\text{C}}}, \tag{39}$$

which represents the **Curie-Weiss law**. For $T_{\text{C}} = 0$ we obtain the **Curie law** for the susceptibility of a paramagnetic material [see Fig. 17 in Sec. 6.3]. For a finite external field B_{ext} the linear curve in Fig. 14(b) is shifted on the y axis [see Eq. 35]. Therefore, there is an intersection of the two curves, and thus a solution for M, for all temperatures. Several $M(T)$ curves for different B_{ext} are shown in Fig. 15(b) for the case $J = 1/2$. Hence, a ferromagnetic material becomes paramagnetic above T_{C} and an external field induces a magnetization.

6.2 Ferrimagnetic order

Ferrimagnets are substances in which the magnetic moment of some ions in the structural unit cell are antiparallel to the others. The ground state is characterized by a spontaneous magnetization that persists up to the Curie temperature. Ferrimagnetism is found in many ferrites, *i.e.* magnetic oxides of the form MO·Fe_2O_3 with the bivalent metal ion M= Zn, Cd, Fe, Ni, Cu, Co, or Mg. Magnetite (Fe_3O_4 or FeO·Fe_2O_3) is a famous example, in which the Fe^{3+}-moments couple antiparallel to each other, and only the Fe^{2+}-ions contribute to the net magnetization.
In general, a ferrimagnetic material can be described to consist of two sublattices A and B with antiparallel spin alignment with the coupling constant $J_{\text{AB}} < 0$. The coupling within the sublattices is described by J_{AA} and J_{BB}, which can be positive, negative, or zero. In the spirit of the mean-field approximation discussed in the previous section the effective field acting on A and B sites are

$$\vec{B}_A^{\text{eff}} = \vec{B}_{\text{ext}} + \mu_0 \lambda_{AB} \vec{M}_B + \mu_0 \lambda_{AA} \vec{M}_A \tag{40}$$
$$\vec{B}_B^{\text{eff}} = \vec{B}_{\text{ext}} + \mu_0 \lambda_{BA} \vec{M}_A + \mu_0 \lambda_{BB} \vec{M}_B, \tag{41}$$

where λ_{AA}, λ_{BB}, λ_{BA}, and λ_{AB} are the respective molecular field constants. For symmetry reasons $\lambda_{AB} = \lambda_{BA}$. Similar to the ferromagnetic case [Eq. (33)] the temperature dependence of the magnetizations of the sublattices $\vec{M}_{A,B}(T)$ are given by Brillouin functions. Above the Curie temperature and for small magnetizations we approximate $B_J(y) \approx \frac{J+1}{3J} y$ and obtain

$$\vec{M}_A = \frac{C_A}{\mu_0 T}(\vec{B}_{\text{ext}} + \mu_0 \lambda_{AA} \vec{M}_A + \mu_0 \lambda_{AB} \vec{M}_B) \tag{42}$$
$$\vec{M}_B = \frac{C_B}{\mu_0 T}(\vec{B}_{\text{ext}} + \mu_0 \lambda_{AB} \vec{M}_A + \mu_0 \lambda_{BB} \vec{M}_B) \tag{43}$$

with the Curie constants C_A and C_B [see Eq. (36)]. This linear set of equations can be solved for $\vec{M}_A$ and $\vec{M}_B$ to obtain an expression for the susceptibility χ and for the ferrimagnetic Curie

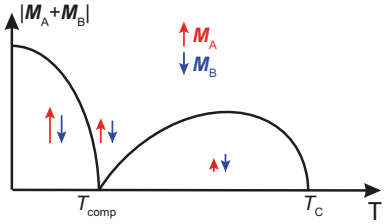


Fig. 16. *Schematic temperature dependence of the total magnetization of a ferrimagnet with a compensation temperature T_{comp}, where the spontaneous magnetizations of the two sublattices $\vec{M}_A$ and $\vec{M}_B$ compensate each other.*

temperature T_C from the condition $\chi^{-1}(T_\mathrm{C}) = 0$. Here, we state for simplicity only the results for $\lambda_{AA} = \lambda_{BB} = 0$ (in many systems this is equivalent to the nearest-neighbor approximation):

$$T_\mathrm{C} = |\lambda_{AB}|\sqrt{C_A C_B} \tag{44}$$

and

$$\chi = \mu_0 \frac{|\vec{M}_A + \vec{M}_B|}{B_\mathrm{ext}} = \frac{(C_A + C_B)T - 2|\lambda_{AB}|C_A C_B}{T^2 - T_\mathrm{C}^2}. \tag{45}$$

Obviously, the temperature dependence of the susceptibility of a ferrimagnet is different from that of a ferromagnet, which allows to experimentally distinguish ferro- and ferrimagnetic materials, see Fig. 17 in Sec. 6.3.

In the general case, the temperature dependence of the total magnetization of a ferrimagnet can show rather complicated behavior depending on the molecular field constants λ_{AB}, λ_{AA}, and λ_{BB}. In particular, the total magnetization can vanish and reappear at a temperature T_comp below T_C due to the different magnitudes and temperature-dependencies of the antiparallel aligned spontaneous magnetizations $\vec{M}_A$ and $\vec{M}_B$ of the sublattices as schematically depicted in Fig. 16. T_comp is called **compensation temperature**.

6.3 Antiferromagnetic order

In an antiferromagnet the exchange coupling between neighboring atoms is negative and the magnetic moments are antiparallel aligned, similar to the situation in a ferrimagnet. However, in an antiferromagnet the two sublattices are identical and their magnetic moments compensate each other. This order persists up to the critical temperature, called **Néel temperature** T_N. An antiferromagnet does not show spontaneous magnetization at any temperature. Hence, antiferromagnetic order cannot be characterized by magnetization measurements, but shows up in neutron scattering as additional diffraction spots due to the twice as large magnetic unit cell. In some antiferromagnetic materials (*e.g.* Cr) the magnitude of the magnetic moments at the atomic sites is modulated with a wavelength longer than the structural periodicity. This is called **incommensurable antiferromagnetic order**. Here, we focus on antiferromagnets with commensurable order such as MnO. From a mean-field point of view, antiferromagnetism of this form is a special case of ferrimagnetism, for which $\vec{M}_A = -\vec{M}_B$, $\lambda_{AA} = \lambda_{BB}$ and hence $C_A = C_B = C$. The effective fields are

$$\vec{B}_A^\mathrm{eff} = \vec{B}_\mathrm{ext} + \mu_0(\lambda_{AB} - \lambda_{AA})\vec{M}_B \tag{46}$$

$$\vec{B}_B^\mathrm{eff} = \vec{B}_\mathrm{ext} + \mu_0(\lambda_{AB} - \lambda_{AA})\vec{M}_A \tag{47}$$

and correspond to the effective fields of a ferrimagnet with renormalized molecular fields constants $\lambda_{AB} \to (\lambda_{AB} - \lambda_{AA})$ and $\lambda_{AA,BB} \to 0$ [see Eqs. (40) and (41)]. Thus, the Néel tempera-

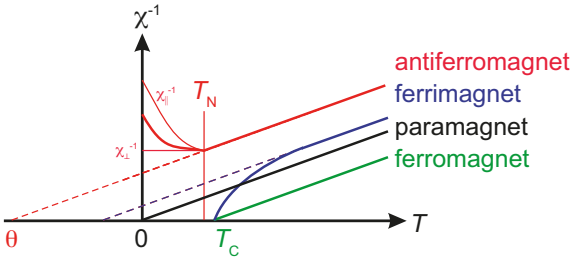


Fig. 17. *Schematic temperature dependence of the inverse susceptibility for para-, ferro-, ferri-, and antiferromagnetic substances. The Curie constants C are assumed the same in all cases leading to the same slope for large T. For $T < T_N$ the susceptibility for a polycrystalline antiferromagnet is the average $\frac{2}{3}\chi_\perp + \frac{1}{3}\chi_\parallel$.*

ture according to Eq. (44) is

$$T_\mathrm{N} = |\lambda_{AB} - \lambda_{AA}|C. \tag{48}$$

Our above definition of C refers to only one sublattice. In the case that C is given for the whole lattice, a factor $\frac{1}{2}$ must be added on the right-hand side.

For $T > T_\mathrm{N}$ in the paramagnetic region, an external field $\vec{B}_\mathrm{ext}$ induces the sublattice magnetizations $\vec{M}_A = \vec{M}_B$ and the total magnetization is $M = 2M_A$. From Eq. (40) follows now

$$\vec{B}_A^\mathrm{eff} = \vec{B}_\mathrm{ext} + \mu_0(\lambda_{AB} + \lambda_{AA})\vec{M}_A = \vec{B}_B^\mathrm{eff} = \vec{B}^\mathrm{eff} \tag{49}$$

and from Eq. (42)

$$\vec{M} = 2\vec{M}_A = \frac{1}{\mu_0}\frac{2C}{T}\vec{B}^\mathrm{eff} = \frac{1}{\mu_0}\frac{2C}{T}\left(\vec{B}_\mathrm{ext} + \mu_0(\lambda_{AB} + \lambda_{AA})\vec{M}_A\right). \tag{50}$$

Finally, we obtain the **susceptibility of an antiferromagnet**

$$\chi = \mu_0\frac{M}{B_\mathrm{ext}} = \frac{2C}{T+\Theta} \quad \text{with} \quad \Theta = -|\lambda_{AB} + \lambda_{AA}|C, \tag{51}$$

where Θ is the **paramagnetic Néel temperature**. The magnitude of Θ differs from T_N for $\lambda_{AA} \neq 0$ and allows the determination of the sign of λ_{AA}. The susceptibility below T_N depends on the direction of $\vec{B}_\mathrm{ext}$ with respect to the spin direction in the sublattices. If it is applied perpendicular to the spins, both sublattice magnetizations rotate slightly into the field direction and $\chi_\perp$ is finite and independent of T. If the field is applied parallel to the spin axis, the magnetic energy does not change and $\chi_\parallel = 0$ at $T = 0$. The qualitative temperature dependence of $\chi_\perp$ and $\chi_\parallel$ is shown in Fig. 17, where we schematically compare the temperature dependence of the inverse susceptibility for para-, ferro-, ferri-, and antiferromagnetic substances.

7 Magnetic anisotropy

Magnetic anisotropy describes the fact that the energy of a magnetic system changes as a function of the direction of the magnetization with respect to the crystal axes or the geometric axes defining the shape of the magnet. Magnetization directions, for which the energy is maximum (minimum) are called **easy (hard) axes of magnetization**. This is manifested in each permanent magnet, where the macroscopic magnetization is pointing up or down with respect to a certain axis, the easy axis, *e.g.* the cylinder axis of a bar magnet or the long axis of a compass

needle. Although the anisotropy energy per atom is much smaller than exchange or bonding energies, almost all applications of magnetic materials depend on the magnetic anisotropy that couples the for humans imperceptible quantity *magnetization* to directly observable orientations and directions in space. Without magnetic anisotropy the magnetization of the compass needle would point to the north pole but not the needle itself, the fridge magnet would not stick to the fridge, and information could not be stably stored in magnetic materials (*e.g.* in hard disk drives).

7.1 Phenomenology of magnetic anisotropy

Phenomenologically the dependence of the free energy density F of a magnetic system on the direction of $\vec{M}$ is expanded in terms of the azimuthal and polar angles θ and ϕ of the magnetization direction. Usually, this is done in terms of the directional cosines $(\alpha_1, \alpha_2, \alpha_3) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$ of $\vec{M}$ with respect to crystal or sample axes. In the absence of magnetic fields time-inversion symmetry requires $F(\vec{M}) = F(-\vec{M})$, such that no odd powers of the cosines appear in the expansion

$$F(\vec{M}) = K_0 + \sum_{ij} b_{ij}\alpha_i\alpha_j + \sum_{ijkl} b_{ijkl}\alpha_i\alpha_j\alpha_k\alpha_l + \ldots, \tag{52}$$

where the coefficients b depend on $|\vec{M}|$ and in general on the temperature. Higher orders in the α's can usually be neglected. Crystal symmetry can further reduce the number of non-vanishing terms. For a cubic crystal Eq. (52) reduces to

$$F(\vec{M}) = K_0 + K_1(\alpha_1^2\alpha_2^2 + \alpha_1^2\alpha_3^2 + \alpha_2^2\alpha_3^2) + K_2\alpha_1^2\alpha_2^2\alpha_3^2 + \ldots \tag{53}$$

Examples of free energy surfaces of the cubic systems bcc-Fe and fcc-Ni are shown in Figs. 18(a) and (b). For $K_1 > 0$ ($K_1 < 0$) [100] directions are easy(hard) axes and [111] directions hard(easy) axes. In both cases [110] directions are intermediate axes. For uniaxial systems like tetragonal or hexagonal lattices, where one axis (usually the c-axis aligned with the $\hat{z}$ direction) is non-equivalent to the other axes, Eq. (52) becomes

$$F(\vec{M}) = K_0 + K_1\sin^2\theta + K_2\sin^4\theta + \ldots \tag{54}$$

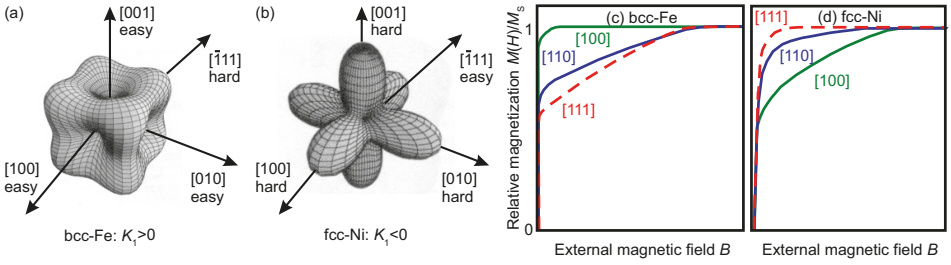


Fig. 18: *Magnetic energy surfaces for cubic symmetry according to Eq. (53) for (a) dominant $K_1 > 0$ representative for bcc-Fe and (b) dominant $K_1 < 0$ representative for fcc-Ni. (c,d) $M(B)$ loops measured with the field B applied along the high-symmetry axes [100], [110], and [111] for (c) bcc-Fe with $K_1 > 0$ and (d) fcc-Ni with $K_1 < 0$.*

As we will see below, the anisotropy constants K_1 and K_2 are typically of the order of 10^4 to 10^7 J/m^3. With an atom density of about 10^{23} cm^{-3} this corresponds to only 1 μeV to 1 meV per atom. The anisotropy constants can be obtained by measuring magnetization loops $M(B)$ with the field applied in different directions with respect to the crystallographic axes or the sample geometry. The area under the $M(B)$ curve between zero field and the saturation field corresponds to the energy required to align the magnetization in the direction of the applied field [Figs. 18(c) and (d)]. Reaching saturation along a hard axis requires a larger field and, thus, a higher energy. For instance, in a cubic system with $K_2 = 0$ we obtain from Eq. (53) for a [100] direction ($\alpha_1 = 1, \alpha_2 = \alpha_3 = 0$) $F_{[100]} = K_0$ and for a [111] direction ($\alpha_1 = \alpha_2 = \alpha_3 = 1/\sqrt{3}$) $F_{[111]} = K_0 + K_1$, hence $K_1 = F_{[111]} - F_{[100]}$.

7.2 Physical origin of magnetic anisotropy

There are different physical mechanisms that give rise to magnetic anisotropy. The most important microscopic mechanisms are discussed below.

7.2.1 Shape anisotropy

A geometrically finite and homogeneously magnetized sample (*i.e.* without magnetic domains) creates a stray field in the outside, which can be thought to be due to magnetic charges (north and south poles) at the surfaces. These magnetic charges also give rise to a dipolar field inside the sample opposite to the magnetization, which is called **demagnetization field** and can lead to a lower energy for certain directions of the magnetization with respect to the sample geometry (shape). Hence, the origin of this so-called shape anisotropy is classical dipole-dipole interaction (Sec. 3.1). Although dipole-dipole interaction is much weaker than exchange, shape anisotropy can be a significant contribution to the total anisotropy, because the weak pair interaction is counterbalanced by the much longer interaction range, which involves a large number atoms.
The energy of a sample with saturation magnetization $\vec{M}_\mathrm{S}$ in its own demagnetization field $\vec{H}_\mathrm{demag}$ is

$$E_\mathrm{stray} = -\frac{1}{2}\int_V \mu_0 \vec{M}_\mathrm{S} \cdot \vec{H}_\mathrm{demag}\, dV, \tag{55}$$

where the integration runs over the whole sample volume V. The calculation of $\vec{H}_\mathrm{demag}$ and the evaluation of the integral for a general shape is not possible in a closed form. However, it can be shown that ellipsoids possess a constant $\vec{H}_\mathrm{demag}$ given by

$$\vec{H}_\mathrm{demag} = -\mathcal{N}\vec{M}_\mathrm{S}, \tag{56}$$

where $\mathcal{N}$ is the **demagnetization tensor**. $\mathcal{N}$ is diagonal in the coordinate system spanned by the semi-axes of the ellipsoid. For a general magnetization direction characterized by the directional cosines in this specific coordinate system

$$\mathcal{N} = \begin{pmatrix} N_1 & 0 & 0 \\ 0 & N_2 & 0 \\ 0 & 0 & N_3 \end{pmatrix} \quad \text{with} \quad N_1 + N_2 + N_3 = 1. \tag{57}$$

The stray field energy for ellipsoids becomes

$$E_\mathrm{stray} = -\frac{1}{2}\int_V \mu_0 \vec{M}_\mathrm{S}\mathcal{N}\vec{M}_\mathrm{S} dV = \frac{\mu_0 V}{2}\vec{M}_\mathrm{S}\mathcal{N}\vec{M}_\mathrm{S} = \frac{\mu_0 V}{2}\cdot M_\mathrm{S}^2(N_1\alpha_1^2 + N_2\alpha_2^2 + N_3\alpha_3^2). \tag{58}$$

For a spherically shaped sample $N_1 = N_2 = N_3 = 1/3$, and $E_{\text{stray}} = \frac{\mu_0 V}{6} M_S^2$ is isotropic. All directions are equivalent. For an infinitely long cylinder in $\hat{z}$-direction, which is a good approximation for a magnetic wire, $N_1 = N_2 = 1/2$ and $N_3 = 0$, and $E_{\text{stray}} = \frac{\mu_0 V}{4} M_S^2 \sin^2\theta$. In this case the cylinder axis ($\theta = 0, \pi$) is an easy axis of the magnetization. This can be understood by the fact that the magnetic charges at *wire ends* are pushed infinitely far apart and the dipolar demagnetization field ($\vec{H}_{\text{demag}} \propto r^{-2}$) approaches zero. Finally, we consider an in the x-y-plane infinitely extended, but thin plate as an approximation for a thin magnetic film. In this case $N_1 = N_2 = 0$ and $N_3 = 1$, and

$$F_{\text{stray}}^{\text{film}}(\vec{M}) = \frac{E_{\text{stray}}^{\text{film}}}{V} = \frac{\mu_0}{2} M_S^2 \cos^2\theta = K_0 + K_{\text{shape}}^{\text{film}} \sin^2\theta \quad \text{with} \quad K_{\text{shape}}^{\text{film}} = -\frac{\mu_0}{2} M_S^2. \tag{59}$$

The magnetization is preferentially in the plane of the film ($\theta = \pi/2, 3\pi/2$), which is called an **easy plane of magnetization**. In order to rotate $\vec{M}_S$ from in-plane to the hard-axis out-of-plane orientation a field $B = \frac{1}{2}\mu_0 M_S$ must be applied, which for Fe, Co, and Ni is about 1.1, 0.9, and 0.3 T, respectively. $K_{\text{shape}}^{\text{film}}$ is the energy density of the shape anisotropy of a thin film and amounts to $1.92 \cdot 10^6$, $1.34 \cdot 10^6$, and $1.73 \cdot 10^5$ J/m^3 for Fe, Co, and Ni, respectively. We will see in the next section that these values are comparable to those of the magnetocrystalline anisotropy and, therefore, the shape anisotropy has a strong impact on the magnetization state of thin films.

7.2.2 Magnetocrystalline anisotropy

Magnetocrystalline anisotropy is a consequence of spin-orbit coupling (SOC) discussed in Sec. 5. The Heisenberg Hamiltonian [Eq. (12)] is isotropic yielding the same energy for all directions of the magnetization. This isotropy is broken when a SOC term is included in the Hamiltonian. This is best seen for the SOC Hamiltonian in Eq. (25), which couples the spin $\vec{S}$ with the angular momentum $\vec{L}$. The alignment of $\vec{L}$ with respect to the crystal lattice, in turn, results from exchange and Coulomb interaction: Atoms with $L \neq 0$ have partly filled orbitals (*e.g.* $3d$-electrons in transition metals or $4f$-electrons in rare-earth metals) resulting in a non-spheric electron distribution. In a crystal the overlap of these non-spheric atomic electron distributions depends on their alignment and, thus, on the direction of $\vec{L}$ with respect to the crystal lattice as schematically shown in Fig. 19. The total effect is a dependence of the free energy density of the magnetic system on the direction of the magnetization with respect to the crystallographic axes as phenomenologically introduced in Sec. 7.1. The magnetocrystalline anisotropy constants of bulk bcc-Fe, fcc-Ni, and hcp-Co are listed in Table 1.
Both the schematic picture in Fig. 19 and the crucial role of the local potential landscape for SOC suggest a strong dependence of the magnetocrystalline anisotropy on the local symmetry of an atom and on its coordination number. At a surface or an interface the symmetry of the crystal's volume is locally broken and thus gives rise to an additional magnetocrystalline

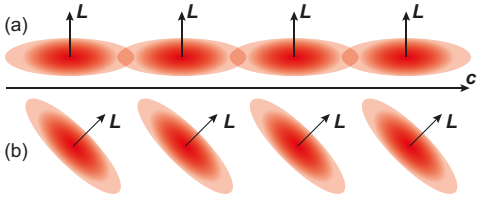


Fig. 19. *Different alignments of the non-spheric atomic electron distributions and the related orbital momenta $\vec{L}$ with respect to a given crystallographic direction $\hat{c}$ result in different interatomic overlap and, hence, different exchange and Coulomb interaction.*

Table 1: *Magnetocrystalline anisotropy constants K_1 and K_2 in J/m^3 of the free energy density expansion Eq. (53) for bulk bcc-Fe, fcc-Ni, and Eq. (54) for hcp-Co, respectively.*

	bcc-Fe	fcc-Ni	hcp-Co
K_1	$5.48 \cdot 10^4$	$-12.63 \cdot 10^4$	$7.66 \cdot 10^5$
K_2	$1.96 \cdot 10^2$	$5.78 \cdot 10^4$	$1.05 \cdot 10^5$

anisotropy, the so-called **interface (or surface) anisotropy**. This interface-induced anisotropy may prefer a magnetization direction in the sample plane or perpendicular to it. The latter case is of particular interest for thin films, because then the interface anisotropy counteracts shape anisotropy. Phenomenologically, interface anisotropy is described by an additional term to the free energy density $F(\vec{M})$

$$F_{\text{interface}}(\vec{M}) = \frac{K_\text{S}}{d} \cdot \sin^2\theta, \tag{60}$$

with the interface/surface anisotropy constant K_S. The film thickness d appears in the denominator because we are dealing with a purely interfacial effect. Note that both interfaces of a thin film may contribute to K_S. The sign and magnitude of K_S depends on the involved materials. Typical values of K_S are $+0.58 \times 10^{-3}$ J/m^2 for a Co/Pd interface, which favors out-of-plane magnetization, or -0.48×10^{-3} J/m^2 for a Ni surface (Ni/UHV interface), which favors in-plane magnetization. Combining shape and interface anisotropy [Eqs. (59) and (60)] for an ultra-thin film we find a spin reorientation transition at the critical thickness d_c, where the total anisotropy crosses zero,

$$d_\text{c} = \frac{2K_\text{S}}{\mu_0 M_\text{S}^2}. \tag{61}$$

Figure 20 shows the experimentally determined total anisotropy constant K of a Co/Pd multilayer multiplied with the Co thickness d as a function of d. The sign change at $d_\text{c} \approx 13$ Å indicates a transition from out-of-plane for small d to in-plane magnetization for large d in good agreement with the prediction of Eq. (61).

Any other symmetry-breaking phenomenon gives rise to additional contributions to the magnetic anisotropy. Regularly aligned step edges on a slightly miscut surface cause **step anisotropy**, which favors certain in-plane directions over others. Strain, for instance due to a mismatch between the lattice constants of the thin film material and the substrate results in **strain anisotropy**, which is closely related to magnetostriction. All types of magnetocrystalline anisotropy are

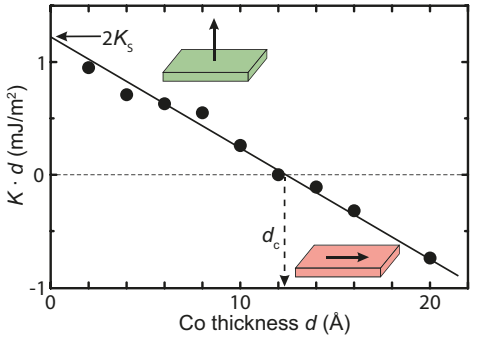


Fig. 20. *Magnetic anisotropy K of thin Co films in a Co/Pd multilayer. The y-intercept of the curve $K \cdot d$* versus *d yields the interface anisotropy constant $2K_S \approx 1.16 \times 10^{-3}$ J/m^3 and the slope the shape anisotropy constant $K_{shape}^{film} \approx -0.91 \times 10^6$ J/m^3. The magnetization is out-of-plane for $d < d_c$ and in-plane $d > d_c$ (data taken from [20]).*

most distinct in single-crystalline materials and are suppressed in polycrystalline specimen without preferred orientation of the grains or in amorphous materials. In these cases shape anisotropy dominates. Magnetocrystalline anisotropy is not restricted to ferro- and ferrimagnetic materials, it also exist in antiferromagnets. In this case, the free energy is a function of the axis defined by the antiferromagnetically coupled sub-lattice magnetizations.

7.2.3 Exchange anisotropy (Exchange biasing)

Another type of anisotropy, which can also be classified as interface anisotropy, is the so-called **exchange anisotropy** occurring at ferromagnet/antiferromagnet (FM/AFM) interfaces. The exchange interaction between the FM and the interface moments of the AFM can be described by an *effective* magnetic field $\vec{H}_{\mathrm{EB}}$ called **exchange-bias field** acting on the FM,

$$\vec{H}_{\mathrm{EB}} = \frac{\sigma_{\mathrm{EB}}}{\mu_0 M d}\hat{e}_{\mathrm{AFM}}, \tag{62}$$

where σ_{EB} is the areal energy density of the exchange coupling across the interface and d the thickness of the FM film. $\hat{e}_{\mathrm{AFM}}$ is the direction of the interface moment of the AFM that arises due to the broken symmetry at the interface and short-range nature of direct exchange with dominant nearest-neighbor and negligible next-nearest-neighbor interaction. The direction $\hat{e}_{\mathrm{AFM}}$ is determined by the magnetocrystalline anisotropy of the AFM material. The corresponding contribution to the free energy density has the form of a Zeeman term

$$F_{\mathrm{EB}}(\vec{M}) = -\mu_0 \vec{M} \cdot \vec{H}_{\mathrm{EB}}. \tag{63}$$

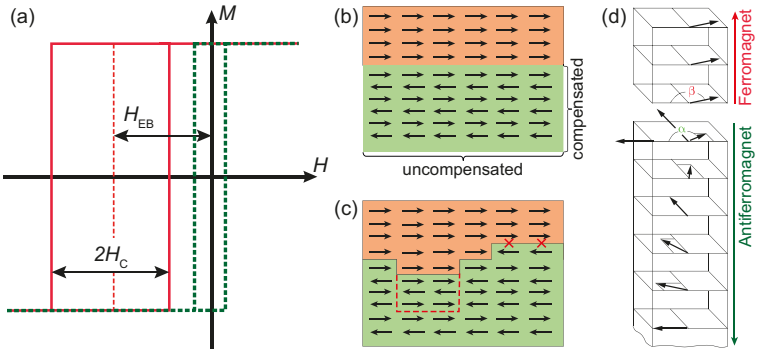


Fig. 21: *(a) Hysteresis loop of a FM film with (red) and without (green) exchange anisotropy due to an adjacent AFM. The loop center of the pinned FM film is shifted to $-H_{EB}$ and the width of the loop $2H_C$ is strongly increased. (b) Simple picture of exchange anisotropy at an ideal interface. All interface spins of the AFM are uncompensated and act in the same way on the FM. A completely compensated surface in marked on the right edge of the AFM. (c) At real interfaces with roughness the nearest-neighbor exchange couplings cannot all be fulfilled simultaneously resulting in frustration (red crosses) and domain walls (dashed red line). (d) If the interfacial exchange coupling is stronger than the anisotropy in the AFM, non-collinear (see angles α and β) configurations are likely and a domain wall in the AFM is formed when the FM film is magnetized to the right. Only one spin sublattice of the AFM is shown in (d) for clarity.*

Therefore, the exchange anisotropy shifts the hysteresis curves of the FM film on the field axis by $-H_{\mathrm{EB}}$, if the field is applied along $\hat{e}_{\mathrm{AFM}}$. This effect is called **exchange biasing** and is schematically shown in Fig. 21(a). The FM is said to be *pinned* by the AFM. The increased width of the hysteresis loop (larger coercive field H_{C}) of the pinned film is a further effect that always comes along with exchange anisotropy and is related to an increased energy needed to reverse the FM when it is coupled to the AFM. In order to properly establish the exchange bias, samples need to be cooled in an applied field from above to below a certain temperature T_{B}, the so-called **blocking temperature**[1] **of exchange bias**. Naturally one would identify T_{B} with the Néel temperature T_{N} of the AFM. However for many systems it is found that exchange bias can only be observed below $T_{\mathrm{B}} < T_{\mathrm{N}}$.

A simplistic and idealized picture of exchange anisotropy is shown in Fig. 21(b), where the AFM is assumed to consist of FM ordered atomic planes stacked with AFM order. If the alignment in the moments in the AFM is rigid (*i.e.* high magnetocrystalline anisotropy), exchange coupling across the interface entering Eq. (62) *via* σ_{EB} is expected to be comparable to direct nearest-neighbor exchange in a FM. For Fe we estimate the areal energy density of direct exchange σ using the Curie temperature $T_{\mathrm{C}} = 1040\,\mathrm{K}$ as a measure for the exchange coupling strength and the lattice constant $a = 2.9\,\text{Å}$, $\sigma \approx k_{\mathrm{B}} T_{\mathrm{C}}/a^2 \approx 170\,\mathrm{mJ/m^2}$. However, experimental values for σ_{EB} are typically less than 1 mJ/m^2 and, thus, more than two orders of magnitude smaller than this estimation. Furthermore, even completely compensated surfaces, where the surface layer contains equal numbers of atoms from both AFM sublattices [*e.g.* the vertical surface in Fig. 21(b)] show an exchange bias effect that cannot be explained in the simple picture sketched above.

Various models have been discussed in order to elucidate these discrepancies. Here, we mention only the **domain-wall model** and the **random-field model**. The domain-wall model [21] attributes the apparent weakening of σ_{EB} to the fact that the moments in the AFM are not completely rigid and are coupled more strongly to those of the FM than to the AFM neighborhood. Thus, domain walls parallel to the interface can form inside the AFM [Fig. 21(d)], and the exchange bias strength is related to the domain wall energy in the AFM, rather than to the interfacial exchange coupling strength. This is different for the random-field model [22], which is based on interface roughness. In the case of an intrinsically uncompensated surface, roughness leads to the formation of terraces with opposite spin direction, and hence to a mesoscopic compensation [Fig. 21(c)]. In the case of an intrinsically compensated surface, however, a small number of uncompensated spins appear due to step edges. In both the compensated and uncompensated case one can assume that planar domains form in the AFM when the FM is ordered in an external field and the system is cooled below T_{B}. In both models, the formation and motion of domain walls in the AFM upon field reversal give rise to the experimentally observed increased coercivity.

In general, the exchange bias effect is thought to be due to uncompensated spins at the surface of the AFM, but their number is much smaller than for an ideal uncompensated surface. A realistic description must include roughness and grain size effects as well as non-collinear spin configurations because any deviation from the ideal situation leads to conflicting interactions: direct exchange between neighbors in the FM and in the AFM as well as across the interface. Frustrated spin configurations and domains are the result. Perpendicular effective interface coupling, where the FM moments are oriented perpendicular to the easy axis of the AFM, is such a frustrated configuration [23]. It has been observed in several systems and demonstrates

[1] This is not to confused with the blocking temperature T_{B} of superparamagnetic nanoparticles and magnetic molecules to be discussed in Sec. 7.3.

the complexity of the exchange bias effect in real samples.

7.3 Superparamagnetism

Magnetic anisotropy plays a crucial role for the stability of the magnetization direction of small magnetized objects against thermal excitation, *e.g.* grains of the magnetic media of hard disk drives (HDD), magnetic elements of magnetic random access memory (MRAM) cells, magnetic nanoparticles, or magnetic molecules. For simplicity we consider magnetic particles with uniaxial magnetic anisotropy in the single domain state. At zero temperature and zero magnetic field the magnetic moments point up or down along the anisotropy axis. The two configurations are energetically equivalent and separated by an energy barrier $\Delta E = K_\mathrm{u} V$, where K_u is the uniaxial anisotropy constant and V the volume of the particles. At finite temperature T and small enough V the thermal energy $k_\mathrm{B}T$ becomes comparable to or larger than ΔE, and thermal excitations can rotate the particles' moments to any direction. The magnetic moments of the particles then behave like the non-interacting atomic moments of a paramagnetic material. Since the particles' magnetization is much larger than atomic moments, the phenomenon is called **superparamagnetism**. Considering a spherical (*i.e.* no shape anisotropy) Co particle and using Eq. (54) we obtain $\Delta E = K_\mathrm{u} V = [F(\theta = \pi/2) - F(\theta = 0)]V = (K_1 + K_2)V$. With the anisotropy constants K_1 and K_2 of Co (Table 1) $\Delta E = k_\mathrm{B}T$ is fulfilled at room temperature for a Co particle with a diameter of 3–4 nm.
The average time between random magnetization reversals due to thermal fluctuations is given by the Néel time

$$\tau_\mathrm{N} = \tau_0 \exp\left(\frac{\Delta E}{k_\mathrm{B}T}\right), \tag{64}$$

where τ_0 is the attempt time, which is typically assumed to be of the order of 10^{-9} s. Obviously, the longer the measurement time τ_m, the higher the probability for a reversal. The **blocking temperature** $T_\mathbf{B}$ is defined as the temperature, for which $\tau_\mathrm{N} = \tau_\mathrm{m}$,

$$T_\mathrm{B} = \frac{\Delta E}{k_\mathrm{B} \ln(\frac{\tau_\mathrm{m}}{\tau_0})}. \tag{65}$$

On the time scale of the measurement, the particle's moment can be considered *blocked* for $T < T_\mathrm{B}$ and thermally excited for $T > T_\mathrm{B}$. For instance, a single atom with an anisotropy energy $\Delta E \approx 1\,\mathrm{meV}$ is blocked for a time of the order of seconds only for $T < 0.5\,\mathrm{K}$. T_B weakly depends on τ_m, but relevant time scales can span several orders of magnitude. Quasi-static magnetic measurements are performed on a time scale of 100 s and thus require according to Eq. (64) a thermal stability factor $\frac{\Delta E}{k_\mathrm{B}T} \approx 25$. The data retention time of a HDD or MRAM of 10 years is guaranteed for $\frac{\Delta E}{k_\mathrm{B}T} > 40$ (*i.e.* $\Delta E > 1\,\mathrm{eV}$ at room temperature), which sets for a material with a given K_u a limit for the minimal volume V of the magnetic grains (HDD) or elements (MRAM). The conflict between miniaturization (minimizing the grain/element volume V), thermal stability (maximizing $K_\mathrm{u}V$), and keeping the required writing field (HDD) or switching current (MRAM), which both increase with increasing K_u, within feasible boundaries is known as the trilemma of magnetic recording and calls for new approaches such as heat-assisted recording, where a short laser pulse is used to reduce the coercivity of the magnetic media during the writing process by heating it locally and temporarily above the Curie temperature.

8 Magnetic domains

In 1907 P. Weiss postulated that a ferromagnet possesses a number of small regions (magnetic domains) in order to explain the response of a magnetic material to an external field. Each domain exhibits the saturation magnetization of the material, but the magnetization direction of different domains are not necessarily parallel. Domains are separated by domain walls.

8.1 Origin of magnetic domains

In an infinite FM crystal the magnetic ground state is characterized by a strict parallel alignment of all magnetic moments along an easy axis. Any deviation costs exchange and anisotropy energy. However, when the crystal is bounded by surfaces (interfaces) magnetic surface charges (poles) give rise to an external stray field [Fig. 22(a)] and thus to a stray field energy contribution. The formation of magnetic domains, *i.e.* different regions with parallel alignment of the moments along different easy-axis directions, can significantly reduce the stray field energy [Fig. 22(b)]. The boundary between domains is not abrupt from one lattice site to the next, which would cost too much exchange energy, but is a smooth transition called domain wall with a continuous rotation of the magnetic moments. Therefore, a domain wall not only costs exchange but also anisotropy energy, which both determine the domain wall energy. The magnetic anisotropy also determines the types of domains that can be formed [Fig. 22(c)].
The domain and domain wall structure established in a sample is a subtle balance between short-range exchange interaction, magnetic anisotropy, and long-range dipolar interaction and is in general rather complex. Domain formation naturally explains the observation that a FM material can exhibit vanishing macroscopic magnetization below the Curie temperature. In fact this demagnetized domain state minimizes the stray field energy. In soft magnetic materials applying a small field of the order of mT or less is sufficient to reach magnetic saturation ($\mu_0 M_S \approx 1\,\mathrm{T}$), because domains rather than individual moments need to be aligned. Domain alignment proceeds *via* growth (shrinking) of the domains aligned along (opposite) to the applied field by domain wall motion, which intrinsically is a reversible low-dissipation process. In real systems crystal defects result in a domain wall potential, which gives rise to **domain wall pinning**, irreversible domain wall motion and enhanced dissipation. Magnetization rotation within the domains comes additionally into play when the external field is applied along a magnetic hard axis. A schematic description of these processes and the resulting magnetization curve is given in Fig. 23. In general the shape of the magnetization loop strongly depends on the domain wall mobility.
Magnetic domains form in all materials that have magnetic ordering due to the exchange in-

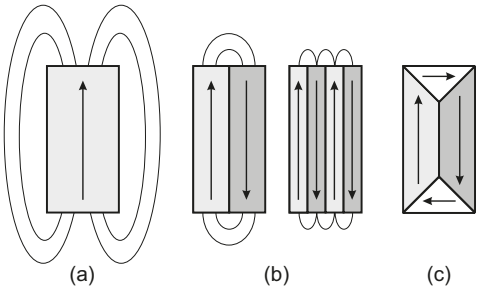


Fig. 22. *(a) Large stray field in the absence of domains. (b) Domain formation reduces the stray field but at the expense of domain wall energy. (c) Magnetic anisotropy influences the domain structure. The Landau domain structure shown in (c) occurs in thin films of materials with triaxial anisotropy,* e.g. *Fe(001) films.*

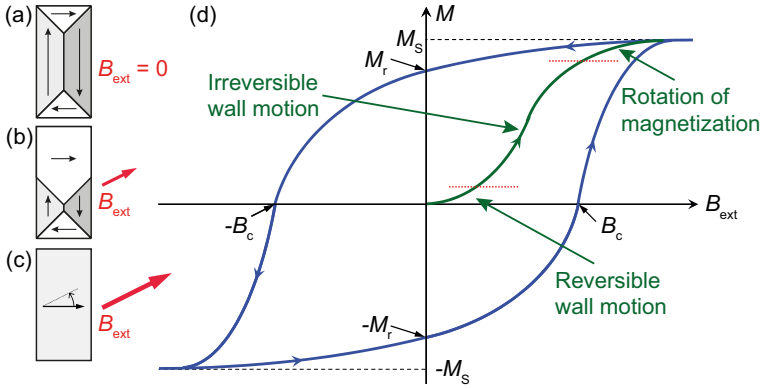


Fig. 23: *Evolution of the domain structure of a single-crystalline, triaxial FM in an external field applied oblique to the easy axes: (a) Landau structure for $\vec{B}_{ext} = 0$, (b) $\vec{B}_{ext}$ in the range of domain wall motions, and (c) in the range of magnetization rotation. (d) Initial magnetization curve (green) and $M(B_{ext})$-loop (blue) with the saturation magnetization M_S, the remanent magnetization M_r, and the coercive field B_c.*

teraction. In addition to ferro- and ferrimagnets, this also includes antiferromagnetic materials. Domains in an antiferromagnet, however, are not the result of the balance between short-range exchange and long-range dipolar interaction, since the vanishing magnetization does not produce a stray field outside of the sample volume. Domains in antiferromagnets are rather stabilized by defects and grains boundaries, which attract and pin domain walls. Therefore, the domain structure in antiferromagnets is strongly dependent on sample history and quality. In contrast to finite ferro- and ferrimagnets, the ground state of an antiferromagnetic specimen is the single-domain state and any decomposition into domains is metastable.

8.2 Domain walls

Domain walls can be classified by the angle between the magnetization directions in the domains separated by the domain wall. Examples of 180° and 90°-walls are shown in Figs. 22(b) and (c). The technologically more relevant 180°-walls can further be divided into **Bloch and Néel walls**. In Bloch walls the magnetization rotates in the plane of the domain wall [Fig. 24(a)], whereas in Néel walls the rotation occurs in a plane that is perpendicular to the plane of the domain wall [Fig. 24(b)]. The distinction becomes important in thin films with in-plane magnetization, for which Bloch walls force the magnetization to rotate out-of-plane thereby creating a strong stray field [Fig. 24(c)]. Therefore, Néel walls [Fig. 24(d)] are energetically more favorable in this situation.
We consider a 180°-wall in a material with uniaxial anisotropy and assume that the magnetization rotates over a distance of N lattice sites. The exchange energy E_φ of two spin enclosing an angle $\varphi = \pi/N$ is according to Eq. (11) and large enough N

$$E_\varphi = -2J\cos\varphi \approx -2J(1-\frac{\varphi^2}{2}) = -2J\left[1-\frac{1}{2}\left(\frac{\pi}{N}\right)^2\right]. \tag{66}$$

The increase in exchange energy per area due to the domain wall for N lattice sites is $E_{ex} =$

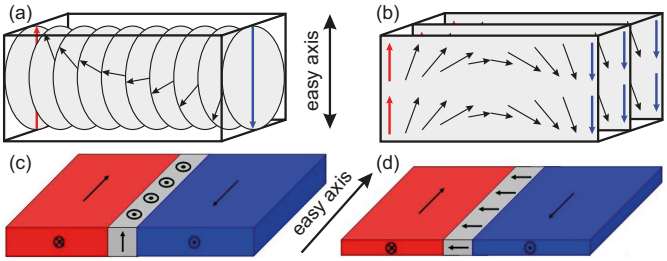


Fig. 24. *Magnetization rotation from an up-domain (red) to a down-domain (blue) in a (a) Bloch and (b) Néel wall. For thin films (c,d) the Néel wall (d) avoids a stray field and is energetically more favorable.*

$J\pi^2/(Na^2)$, where $1/a^2$ is the number of spins per area and a the lattice constant. If the domain wall was atomically sharp, then only one spin pair per unit area a^2 would change its exchange energy from $-2J$ to $+2J$, and we directly see that an domain wall with finite width $N > 2$ is always energetically more favorable. Each lattice site n ($n = 1 \ldots N$) also contribute an anisotropy energy $K_1 a^3 \sin^2(n\pi/N)$ [Eq. (54)] yielding a total anisotropy energy per area of the domain wall

$$E_{\text{ani}} = \sum_{n=1}^{N} K_1 a \sin^2(\frac{n\pi}{N}) \approx K_1 a \frac{N}{\pi} \int_0^{\pi} \sin^2 x \, dx = \frac{1}{2} N K_1 a. \tag{67}$$

The total **domain wall energy** $E_{\text{ex}} + E_{\text{ani}}$ is minimized for $N_0 = \pi\sqrt{2J/(K_1 a^3)}$ yielding the **domain wall width** d_{dw} and the domain wall energy per area E_{dw}

$$d_{\text{dw}} = N_0 a = \pi\sqrt{\frac{A}{K_1}} \quad \text{and} \quad E_{\text{dw}} = \pi\sqrt{AK_1}, \tag{68}$$

where $A = 2J/a$ is the so-called **exchange stiffness**. A strong exchange stiffness favors wide domain walls, whereas strong anisotropy reduces the width. For Fe we find $d_{\text{dw}} \approx 40\,\text{nm}$.
The above estimations are valid for both Bloch and Néel walls in the bulk of a crystal. For thin films, however, the stray field energy outside the film or in other words shape anisotropy according to Sec. 7.2.1 must additionally be taken into account. Whereas Néel walls avoid stray fields [Figs. 24(c) and (d)], Bloch walls experience an additional shape anisotropy energy density of the form $M_S^2/2 \cdot \sin^2\varphi$, where φ is measured from the easy axis in the plane of the film. Hence, K_1 in Eq. (68) must be replaced by $K_1 + M_S^2/2$ making Bloch walls in thin films energetically less favorable than Néel walls.

9 Electrical transport in magnetic metals

Electrical transport in metals reflects the character of conduction electrons at the Fermi level. Metals with high conductivity usually have conduction electrons of s or p-character, whereas f-electrons hardly contribute to transport. d-electrons are somewhere in between. In transition metals the d-electrons play an important role and lead to a connection between magnetism and transport properties. In this chapter some key concepts of electrical transport in magnetic metals, like **Mott's two-channel model, spin polarization, and spin accumulation** will be introduced. The technologically most important spin-transport phenomena **anisotropic, giant, and tunneling magnetoresistance (AMR, GMR, and TMR)** as well as the manipulation of

magnetization by **spin-transfer torques (STT)** will be discussed in the following Secs. 10, 11, 12, and 13.

9.1 Boltzmann equation and relaxation time approximation

First we consider a non-magnetic metal. At thermal equilibrium, the single particle energy levels are occupied according to the **Fermi-Dirac equilibrium function**

$$f_0(\epsilon(\vec{k})) = \frac{1}{1+\exp[(\epsilon(\vec{k})-\epsilon_\mathrm{F})/k_\mathrm{B}T]}. \tag{69}$$

As a consequence only few electrons with an energy close enough to the Fermi level ϵ_F can contribute to electrical transport, all others are hindered by the Pauli principle. The conductivity is the result of changes of the distribution function $f(\vec{r},\vec{k},t)$ due to (i) external forces (*e.g.* electric or magnetic fields $\vec{E}$ and $\vec{B}$), (ii) diffusion caused by spatial gradients of the electron density, and (iii) dissipation resulting from scattering processes as described by the **Boltzmann equation**

$$\left(\frac{\partial f}{\partial t}\right)_\mathrm{total} = -\frac{e}{\hbar}(\vec{E}+\vec{v}\times\vec{B})\cdot\nabla_{\vec{k}}f - \vec{v}\cdot\nabla_{\vec{r}}f + \left(\frac{\partial f}{\partial t}\right)_\mathrm{scatt}. \tag{70}$$

Usually the deviations from the thermal equilibrium are small and the Boltzmann equation can be linearized. As a further simplification the scattering term is expressed using the **relaxation time approximation**

$$\left(\frac{\partial f}{\partial t}\right)_\mathrm{scatt} = -\frac{f(\vec{k})-f_0(\vec{k})}{\tau(\vec{k})}, \tag{71}$$

where τ is the relaxation time. It determines the rate of return to equilibrium and therefore is a measure for the scattering strength. Under these assumptions the current density $\vec{J}=\hat{\boldsymbol{\sigma}}\cdot\vec{E}$ can be calculated, where $\hat{\boldsymbol{\sigma}}$ is the **conductivity tensor**

$$\hat{\boldsymbol{\sigma}} = \frac{e^2}{4\pi^3\hbar}\int_\mathrm{FS}\frac{\tau(\vec{k})\,\vec{v}(\vec{k})\cdot\vec{v}(\vec{k})}{v(\vec{k})}\,d\vec{S}. \tag{72}$$

The integration runs over the Fermi surface. For isotropic or cubic materials the conductivity is a scalar and using for simplicity a Fermi sphere for the $\vec{k}$-space integration (free-electron model), we obtain

$$\sigma = \frac{ne^2\tau}{m_\mathrm{eff}}, \tag{73}$$

where n is the carrier density, e the elementary charge, and $m_\mathrm{eff}=\hbar^2(\partial^2\epsilon/\partial k^2)^{-1}$ the effective mass.[2] The number of carriers at ϵ_F contributing to transport and the effective mass are tightly connected to the bandstructure explaining the above mentioned differences between s, p, d, and f-electrons, whereas the scattering processes enter *via* τ. For metals with a perfect periodic lattice structure and at $T=0$ the eigenfunctions of the Schrödinger equation are Bloch waves with infinite relaxation time (mean free path) and thus zero resistance. Defects and lattice vibrations (phonons) as well as electron-electron interaction give rise to incoherent scattering

[2] This is precisely the result of the classical Drude model, where *all* electrons (density $\propto k_\mathrm{F}^3$) move with the small drift velocity due to $\vec{E}$ and contribute to transport. Here, however, only few electrons at ϵ_F (density $\propto k_\mathrm{F}^2$) but moving with the drift velocity added the much larger Fermi velocity $v_\mathrm{F}=\hbar k_\mathrm{F}/m$ are involved.

of the Bloch states and hence resistivity. A detailed description of the scattering mechanisms is beyond the scope of this lecture. Each scattering mechanism is characterized by a relaxation time τ_i, and according to the **Matthiessen rule** the total relaxation time is given by $\tau^{-1} = \sum_i \tau_i^{-1}$.

9.2 Normal and spin-disorder magnetoresistance

Magnetoresistance describes any dependence of the resistance on the magnetic field $\vec{B}$ acting on the sample. The total field $\vec{B}$ comprises contributions from the external field $\vec{B}_{\text{ext}}$, the demagnetizing field $\mu_0 \vec{H}_{\text{demag}}$, and the field of sample magnetization $\mu_0 \vec{M}$. Hence, magnetoresistance also includes history-dependent, hysteretic effects of the (remanent) magnetization state, which can be controlled by applying external magnetic fields in different directions or sequences, on the resistance.

The **normal or positive magnetoresistance** is a consequence of the Lorentz force [first term in Eq. (70)] of an applied field acting on moving charge carriers. Under this force, the carriers move between two scattering events on circular trajectories. This leads to a reduction of the effective mean free path l and equivalently the relaxation time $\tau = l/v_{\text{F}}$ resulting in the Boltzmann formalism in an enhanced resistivity. The resistivity increases with the magnetic field strength, and thus the normal magnetoresistance is classified as a *positive magnetoresistance*. The term *normal* refers to the fact that this kind of magnetoresistance occurs in all conductive materials no matter if they are magnetic or not. The relative resistivity chance due to the applied field $\Delta\rho/\rho_0$ follows the **Kohler rule**

$$\frac{\Delta\rho}{\rho_0} = \frac{\rho(B) - \rho(0)}{\rho(0)} = F\left(\frac{B}{\rho_0}\right), \tag{74}$$

where $F(x)$ is a function that depends on the type of metal. The normal magnetoresistance can be large for extremely clean and perfect metals with very low ρ_0, *e.g.* up to 5% in Cu or Ag at low temperatures and 10 T. At room temperature however, where ρ_0 is enhanced, the effect is in general very small and not usable for applications.

For ferromagnetic transition metals an externally applied field has an additional and opposite effect on the resistivity. Below the Curie temperature T_{C} ferromagnetically ordered metals (*e.g.* Fe, Co, Ni) have a *lower* resistance than non-ferromagnetic transition metals with similar electronic structure (*e.g.* Pd). This *negative* magnetoresistance can be related to the exchange-splitting of the bandstructure that is associated with ferromagnetic order in a metal (Sec. 4). The $3d$-states of the majority DOS drop below the Fermi level and electrons cannot be scattered into these states anymore resulting in a reduced resistivity in the ferromagnetically ordered state. The temperature dependence of the negative magnetoresistance shows a increase towards the Curie temperature T_{C} that cannot be explained by the exchange-split DOS alone, since the exchange-splitting is progressively reduced by thermally induced spin disorder. Spin disorder, however, gives rise to inelastic scattering from spinwaves (magnons). At a given temperature below T_{C} an external magnetic field counteracts the thermally induced fluctuations of the magnetic moments and increases the spin order thereby reducing the spin-disorder scattering and the resistivity, hence the name **spin-disorder magnetoresistance**. At typical operation temperatures well below T_{C} and acceptable magnetic fields the spin-disorder magnetoresistance is too small for applications.

9.3 Spin polarization and spin accumulation

In the description in Sec. 9.1 the spin of the charge carriers did not play a role for transport. In order to describe ferromagnets we consider the two-channel model. Already in 1935 Mott [24] suggested to *formally* split the current into two independent, parallel contributions, the spin-up and spin-down channels. This hypothesis assumes that spin-flip scattering is negligible and that the two channels contribute differently to transport. The latter can be rationalized by considering the key quantities for the expression of the electrical conductivity [Eqs. (72) and (73)]. The carrier density at the Fermi level as well as the effective mass m_{eff}, which is related to the curvature of the band dispersion $\epsilon(\vec{k})$, are in a ferromagnet different for majority and minority spins due to the exchange splitting of the band structure, see Figs. 8, 10, and 11. Also the scattering processes and hence τ can be spin-dependent, *e.g.* due to spin-orbit interaction (Sec. 10) or due to the spin-split DOS at the Fermi energy, which provides for the two spin channels different densities of final states available for scattering processes.
A direct consequence of the different conductivities for majority and minority spin channels in a ferromagnet is a polarization of the current defined by

$$P = \frac{J^{\text{maj}} - J^{\text{min}}}{J^{\text{maj}} + J^{\text{min}}} \approx \frac{N^{\text{maj}} - N^{\text{min}}}{N^{\text{maj}} + N^{\text{min}}}, \tag{75}$$

where $J^{\text{maj,min}}$ are the current densities in the two channels. In practice $J^{\text{maj,min}}$ are not known and usually cannot be measured. Hence, P is approximately calculated by replacing $J^{\text{maj,min}}$ by the majority and minority DOS at the Fermi level $N^{\text{maj,min}}$, thereby neglecting the spin dependence of the electron mobility $\mu = |e|\tau/m_{\text{eff}}$. $|P| = 100\%$ indicates a completely polarized current, whereas $P = 0$ corresponds to an unpolarized current, which is the equilibrium situation in a paramagnetic metal (*e.g.* Ag, Au, Cu). If a current flows from a ferromagnet with $P > 0$ ($P < 0$) into a paramagnet without asymmetry between the spin channels, there will be a surplus of majority (minority) electrons in the paramagnet, which induces a small magnetic moment per volume in the paramagnet. This imbalance of the spin distribution deviating from the equilibrium (*i.e.* without current flow) is called **spin accumulation**. Obviously, a constant current applied across a ferromagnet/paramagnet cannot lead to steady increase of the spin accumulation. The reason is **spin-flip scattering** that transfers electrons in states with surplus spin character into states with opposite spin. The result is a dynamic equilibrium between spin injection into the paramagnet and spin relaxation by spin-flip scattering, which induces an average spin accumulation as long as the current is applied.
The characteristic length scale over which the spin accumulation decays in a paramagnetic material is given by the spin diffusion length λ_{sdiff}

$$P(x) = P_0 \exp(-\frac{x}{\lambda_{\text{sdiff}}}), \tag{76}$$

where x is the distance from the ferromagnet/paramagnet interface at $x = 0$ and P_0 the injected spin polarization. Spin-flip scattering occurs with a much lower probability than spin-conserving momentum scattering. An electron must undergo $N \approx 10^3$ collisions before it encounters a spin-flip process after an average time τ_{sf}. In a random-walk model the electron penetrates a distance $x = \lambda_{\text{sdiff}} = l\sqrt{N/3}$ into the paramagnet during this time, where l is the mean free path of momentum scattering, and travels a total distance $Nl = v_{\text{F}}\tau_{\text{sf}}$. Eliminating N from the two expressions yields

$$\lambda_{\text{sdiff}} = \sqrt{\frac{l v_{\text{F}} \tau_{\text{sf}}}{3}}. \tag{77}$$

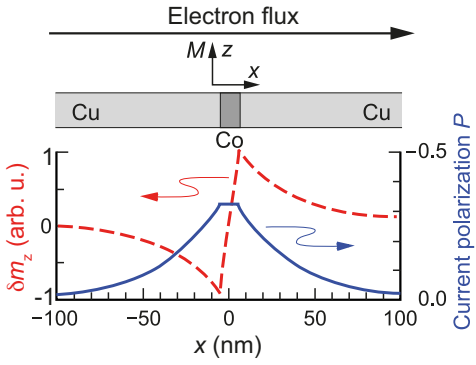


Fig. 25. *Spin accumulation δm_z (red) and current polarization P (blue) due to a current flowing through a ferromagnetic Co layer embedded in semi-infinite Cu leads. After Ref. [25].*

λ_{sdiff} depends on the material and *via* l also on extrinsic properties like crystallinity and purity. Typical values range from a few nanometers (*e.g.* permalloy $Ni_{80}Fe_{20}$) up to several tens of nanometers for magnetic metals (*e.g.* Co) and exceeds 100 nm for non-magnetic metals (*e.g.* Cu).

Figure 25 shows the spin accumulation (red) and current polarization (blue) for a current flowing (electron flux in $+x$-direction) through a thin Co layer magnetized in $+z$-direction embedded in semi-infinite Cu leads. The current polarization in bulk Co is negative, hence minority spin electrons prevail downstream of the Co layer giving rise to a positive spin accumulation δm_z, while majority spin electrons accumulate upstream inducing a negative δm_z. The spin accumulation exponentially decays due to spin-flip scattering with distance $|x|$ from its sources, namely the Co/Cu interfaces, with the characteristic length scale given by λ_{sdiff}. This is reflected in Fig. 25 by the much faster decay of δm_z in Co compared to Cu. Any spatial gradient in the spin accumulation gives rise to a spin-polarized current that counteracts the imbalance in the same manner as a gradient in a particle density induces a diffusive particle current. The surplus of transmitted minority electrons leads to a negative P in the Cu lead downstream of the Co layer. On the upstream side, the reflected flux of majority electrons also leads to an negative polarization of the total (unpolarized incoming plus positively polarized reflected) current.

10 Anisotropic magnetoresistance (AMR)

Anisotropic magnetoresistance (AMR) describes the dependence of the electric resistivity of a FM material on the angle between the current and the magnetization direction. AMR is a volume effect discovered in 1857 and was applied in read heads of hard-disk drives since the 1970s until the implementation of GMR read heads in 1998.

10.1 Phenomenological description

Figure 26 shows an example how the resistivity of a FM changes when an external field is applied either parallel $[\rho_{\parallel}(H)]$ or perpendicular $[\rho_{\perp}(H)]$ to the current direction. In the demagnetized state (domain configuration with no net total magnetization) at zero field there is no difference between the two field orientations. Upon increasing the field strength, $\rho_{\parallel}(H)$ and $\rho_{\perp}(H)$ show different behavior. For most FM materials, $\rho_{\perp}(H)$ decreases and $\rho_{\parallel}(H)$ increases. These initial effects are due to the reorientation of the magnetic domains up to the

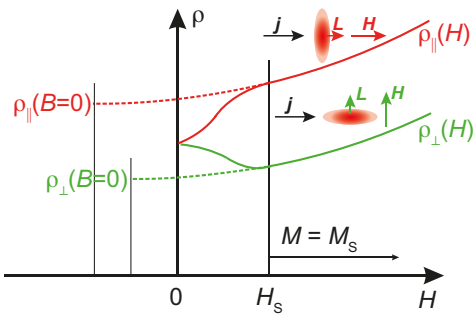


Fig. 26. *Schematic resistivity changes of a FM upon applying an external field parallel $\rho_\parallel(H)$ or perpendicular $\rho_\perp(H)$ to the current direction $\vec{j}$. Dashed lines show the extrapolations to vanishing internal field B for the determination of $\rho_\parallel$ and $\rho_\perp$. The increase of resistivity for $H > H_S$ is due to the normal magnetoresistance. Insets schematically show the different scattering cross-sections for $\vec{j}$ parallel and perpendicular to $\vec{H}$, respectively.*

saturation field H_S. Above H_S the slow and for both field orientations equal evolution of the resistivity with field is due to the normal (positive) and the (negative) spin-disorder magnetoresistance (Sec. 9.2). In Fig. 26 the normal magnetoresistance is dominating. For a quantitative description of AMR the $\rho_\parallel(H)$ and $\rho_\perp(H)$ curves are extrapolated to the fields H, for which the internal fields B vanish to obtain the parameter $\rho_\parallel \equiv \rho_\parallel(B=0)$ and $\rho_\perp \equiv \rho_\perp(B=0)$. The difference between $\rho_\parallel$ and $\rho_\perp$ is called **spontaneous resistivity anisotropy**, which is related to the spontaneous magnetization and thus vanishes above the Curie temperature. The magnitude of the AMR effect is expressed as the ratio between spontaneous resistivity anisotropy and the direction-averaged resistivity

$$\frac{\Delta R}{R} = \frac{\rho_\parallel - \rho_\perp}{\frac{1}{3}\rho_\parallel + \frac{2}{3}\rho_\perp}. \tag{78}$$

The angular dependence of the resistivity due to AMR is

$$\rho(\theta) = \rho_\perp + (\rho_\parallel - \rho_\perp) \cdot \cos^2(\theta), \tag{79}$$

where θ is the angle between the current direction and the magnetization, which for large enough field is parallel to the field direction. The spontaneous resistivity anisotropy ratio reaches at low temperature for alloys like NiFe or NiCo up to 20%, but decreases to a few percents at room temperature. $Ni_{80}Fe_{20}$ (permalloy) combines an AMR effect of up to 3% at room temperature and soft magnetic behavior making it easy to change the magnetization direction with an external field and is therefore frequently used for applications. In general the AMR effect is much smaller for most FM materials, *e.g.* 0.3% for bcc-Fe.

10.2 Microscopic picture: Scattering into spin-orbit-coupled states

The origin of the AMR effect is spin-orbit coupling (SOC) discussed in Sec. 5. SOC results in an orbital contribution to the atomic moment and hence a non-spheric atomic electron distribution as already discussed in the context of magnetocrystalline anisotropy (Sec. 7.2.2). The asymmetry of the charge distribution is connected to the direction of the of the orbital moment and *via* SOC to the direction of the spin moment. A rotation of the spin moment by an external field also rotates the non-spheric electron distribution. The different resistance for a current flowing parallel or perpendicular to the magnetization direction is a consequence of the different scattering cross-sections due to the non-spheric electron distribution as schematically shown in the insets in Fig. 26.

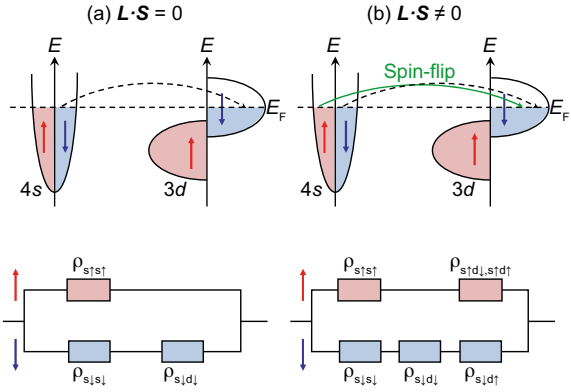


Fig. 27. *Spin-split DOS of a strong 3d-ferromagnet and the spin separated contributions to the resistivity in Mott's two-channel model (a) in the absence of SOC and (b) with SOC included. SOC causes weak $d^{\uparrow}$-$d^{\downarrow}$ mixing and opens up spin-flip transitions. The 3d-DOS represents in (a) eigenstates with* pure *spin-up or spin-down character and in (b), however, eigenstates with* mainly *spin-up or spin-down character with SOC-induced admixtures of opposite spin character, see Eqs. (80) to (84).*

For a more in-depth understanding of AMR we consider in Fig. 27 the schematic DOS of a strong $3d$ FM transition metal, *e.g.* Co, and Mott's two-channel model (Sec. 9.3). The current is mainly carried by s-electrons due to their lower effective mass compared to the d-electrons. We first neglect SOC. The DOS at the Fermi level allows for s-s and s-d scattering transitions, where the s-d transitions contribute most to the resistivity due to the larger DOS of the d-states at $E_{\rm F}$. In the FM phase exchange-splitting leads to a spin-dependent asymmetry: The majority $d^{\uparrow}$-states get completely filled and are no longer available for scattering events. There is only $s^{\uparrow}$-$s^{\uparrow}$ scattering in the majority channel, which is represented in the circuit diagram of Fig. 27(a) by the resistivity $\rho_{s^{\uparrow}s^{\uparrow}}$. In the minority channel, however, scattering of $s^{\downarrow}$-states into $d^{\downarrow}$-states is possible in addition to $s^{\downarrow}$-$s^{\downarrow}$ scattering [Fig. 27(a)]. Note that all these resistivities are isotropic and do not contribute to AMR.

Now we switch on SOC by including the SOC Hamiltonian $\mathcal{H}_{\rm SOC} = \xi\,\vec{L}\cdot\vec{\boldsymbol{S}}$ [Eqs. (25) and (26)] in the Hamiltonian. We have already discussed in Sec. 5 how SOC leads to an admixing of components with opposite spin character in the eigenstates. Following the arguments of Smit [26] and Campbell *et al.* [27] we consider a tight-binding model for the d-states with an exchange field $H_z^{\rm ex}$ but no crystal field. Without SOC the five d-states per spin channel are degenerate and equally occupied and have spatially the form of the atomic d-wavefunctions $\phi|m,s\rangle$ with $m = 0, \pm1, \pm2$ and $s = \uparrow, \downarrow$. Taking the SOC perturbation [Eqs. (25) and (26)] with ξ small compared to $H_z^{\rm ex}$ into account, the wavefunctions with *mainly* spin-down character become [27]

$$\Psi|2\downarrow\rangle = (1-\frac{1}{2}\epsilon^2)\,\phi|2\downarrow\rangle + \epsilon\,\phi|1\uparrow\rangle \tag{80}$$

$$\Psi|1\downarrow\rangle = (1-\frac{3}{4}\epsilon^2)\,\phi|1\downarrow\rangle + \sqrt{\frac{3}{2}}\,\epsilon\,\phi|0\uparrow\rangle \tag{81}$$

$$\Psi|0\downarrow\rangle = (1-\frac{3}{4}\epsilon^2)\,\phi|0\downarrow\rangle + \sqrt{\frac{3}{2}}\,\epsilon\,\phi|-1\uparrow\rangle \tag{82}$$

$$\Psi|-1\downarrow\rangle = (1-\frac{1}{2}\epsilon^2)\,\phi|-1\downarrow\rangle + \epsilon\,\phi|-2\uparrow\rangle \tag{83}$$

$$\Psi|-2\downarrow\rangle = \phi|-2\downarrow\rangle, \tag{84}$$

where $\epsilon = \xi/H_z^{\rm ex}$. This $\boldsymbol{d^{\uparrow}}$**-**$\boldsymbol{d^{\downarrow}}$ **mixing** allows for spin-flip scattering processes and new scatter-

ing channels are opened. $s^{\uparrow}$-states can now scatter into $d^{\downarrow}$-states giving rise to $\rho_{s^{\uparrow}d^{\downarrow}}$. This most important new spin-flip transition is shown as a green arrow in [Fig. 27(b)]. Likewise $d^{\uparrow} \rightarrow s^{\uparrow}$ transitions become possible, thereby creating unoccupied $d^{\uparrow}$-states. These empty states open up further channels for spin-flip and non-spin-flip scattering, *e.g.* $s^{\downarrow} \rightarrow d^{\uparrow}$ [Fig. 27(b)]. The $d^{\uparrow}$-$d^{\downarrow}$ mixing is not isotropic as can be seen form the different coefficients of the spin-up admixtures in Eqs. (80) to (84) because the magnetization direction provides an axis for the spin-orbit perturbation. If we assume that s-electrons are plane waves $\exp(i\vec{k}\cdot\vec{r})$ and that the s-d scattering potential $V(r)$ is spherical, the s-d transition probabilities $|\langle e^{i\vec{k}\cdot\vec{r}}|V(r)|\Psi|m,s\rangle|^2$ (scattering cross-sections) can be calculated for s-electrons with different propagation directions $\vec{k}$. It turns out that, for example, s-states $\exp(ik_z\cdot z)$ propagating along the magnetization direction (given by H_z^{ex}) can only be scattered into d-states with $m=0$ and states $\exp(ik_x\cdot x)$ only into d-states with $m=0,\pm 2$, with appropriate coefficients in each case. The spin-separated resistivities for initial propagation directions k_z and k_x become

$$\rho_{s^{\uparrow}d}(k_z) = \frac{3}{2}\epsilon^2\rho' \quad (85)$$

$$\rho_{s^{\uparrow}d}(k_x) = \frac{3}{4}\epsilon^2\rho' \quad (86)$$

$$\rho_{s^{\downarrow}d}(k_z) = (1-\frac{3}{2}\epsilon^2)\rho' \quad (87)$$

$$\rho_{s^{\downarrow}d}(k_x) = (1-\frac{3}{4}\epsilon^2)\rho', \quad (88)$$

where ρ' is the s-d resistivity for spin-down electrons in the absence of SOC. The result of these simplified considerations shows that upon inclusion of SOC part of the resistivity $\rho_{s^{\downarrow}d}$ is transferred to $\rho_{s^{\uparrow}d}$ and that this effect is twice as strong for electrons traveling along the magnetization (z-direction) as for those traveling perpendicular to it (x-direction). In this way, only s-d transitions enabled by SOC provide a mechanism for the resistivity anisotropy, which is the origin of AMR.

11 Giant magnetoresistance (GMR)

The description of the electrical resistance in ferromagnets (Sec. 9.3) based on Mott's two-channel model suggests that the mobility of conduction electrons in a FM metal depends on the orientation of their spin with respect to the magnetization direction. Testing this suggestion turned out to be an extremely difficult task as there was no experimental means to precisely control the relative spin orientation of spin scatterers. The situation changed completely with the discovery of interlayer exchange coupling (IEC, see Sec. 3.4) in 1986: Multilayers of FM layers separated by nm-thick non-magnetic (NM) spacer layers exhibiting antiferromagnetic IEC provide a means to control the relative alignment of neighboring spin scatterers (*i.e.* the magnetization of neighboring FM layers) within the mean free path of the electrons (*i.e.* the thickness of the spacer layers). Therefore, it is not surprising that GMR was discovered in the wake of IEC.

11.1 Phenomenological description

The giant magnetoresistance (GMR) effect describes the finding that in layered magnetic structures the resistivity depends on the relative alignment of the magnetizations of adjacent FM

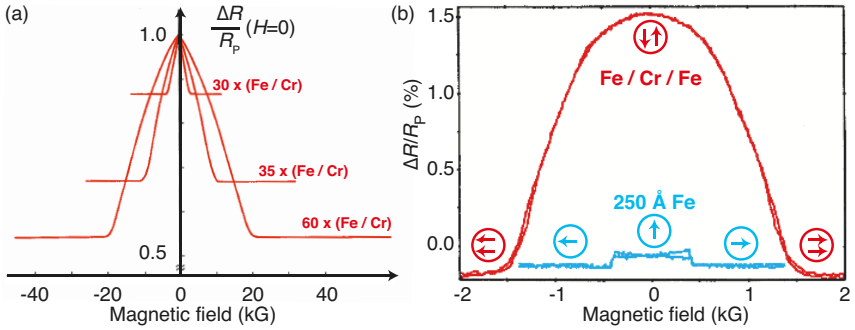


Fig. 28: *First observations of the GMR effect in (a) Fe(3 nm)/Cr(d_{Cr}) multilayer for a out-of-plane applied magnetic field measured at 4.2 K [4] and (b) Fe (12 nm)/Cr(1 nm)/Fe (12 nm) trilayers for an in-plane applied field measured at 300 K [5]. The blue curve in (b) shows the AMR effect of a single Fe layer with a thickness equal to the total thickness of the Fe/Cr/Fe trilayer for comparison.*

layers. The pioneering experiments, both exploiting Fe/Cr layered structures with strong antiferromagnetic IEC (Sec. 3.4), are displayed in Fig. 28. At zero field IEC aligns the magnetizations of adjacent Fe layers antiparallel, whereas a large enough external magnetic field saturates the sample and forces the Fe layers into a parallel configuration. The transition from the antiparallel to the parallel alignment is accompanied by a drastic change of the resistivity. The blue curve in Fig. 28(b) shows the AMR effect of a 250 Å Fe layer for comparison. The much larger response of the layered structures is the reason why the new effect was dubbed *giant* magnetoresistance. The measurements in Fig. 28 represent the simultaneous, but independent discovery of GMR, for which Albert Fert (University of Paris-Sud) and Peter Grünberg (Research Center Jülich) were awarded the Nobel Prize in Physics in 2007.

Apart from antiferromagnetic IEC the antiparallel alignment of the layers' magnetizations at small fields can also be achieved by hysteresis effects. In the latter case one film is magnetically pinned (*e.g.* by the exchange bias effect due to an antiferromagnet as discussed in Sec. 7.2.3), whereas the magnetization of the other is free to rotate when an external field is applied. Such arrangements are called **spin valves** and are relevant for applications. An example of a spin valve is shown in Fig. 29. The steep slope of resistance near zero field provides a sensitive signal to measure small magnetic fields.

If we denote by R_P the resistance for parallel alignment of adjacent ferromagnetic films and by R_AP the same for antiparallel alignment, then the strength of GMR effects is usually quoted in terms of

$$\frac{\Delta R}{R_\mathrm{P}} = \frac{R_\mathrm{P} - R_\mathrm{AP}}{R_\mathrm{P}}. \tag{89}$$

Mostly, the resistance is highest for antiparallel alignment yielding a positive $\Delta R/R_\mathrm{P}$ corresponding to the so-called **normal GMR effect**. But there are also cases, where the situation is reversed and $\Delta R/R_\mathrm{P}$ becomes negative. This is called the **inverse GMR effect**.

The GMR effect has been investigated in two different geometries, namely the CIP (Current-In-Plane) and the CPP (Current-Perpendicular-Plane) geometry. The relative effect is stronger in the CPP geometry. However, due to the extremely unfavorable geometric conditions (lateral dimensions some orders of magnitude larger than the film thickness), the voltage drop per-

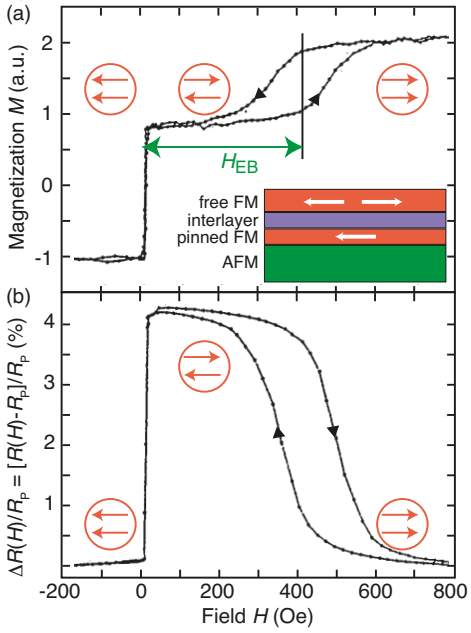


Fig. 29. *Characteristics of a spin valve: The layer sequence of the spin valve is $Fe_{20}Ni_{20}$(6 nm)/Cu(2.2 nm)/$Fe_{20}Ni_{20}$(4 nm)/FeMn (7 nm), where the antiferromagnetic FeMn layer pins the lower $Fe_{20}Ni_{20}$ layer due to the exchange bias effect. (a) Magnetization loop with a sharp switching of the unpinned, free $Fe_{20}Ni_{20}$ layer at zero field and the shifted hysteresis loop of the exchange-biased, pinned $Fe_{20}Ni_{20}$ around H_{EB}. Pairs of orange arrows indicate the relative alignment of the magnetizations of the magnetic films. (b) Magnetoresistance ratio $\Delta R(H)/R_P = [R(H) - R_P]/R_P$ measured at room temperature showing a clearly enhanced resistance for antiparallel alignment. After Ref. [28].*

pendicular to the layers (CPP geometry) is very difficult to detect without special structuring. Typical values for the GMR effect as defined by Eq. (89) both in the CIP and the CPP geometry are of the order of 10% at room temperature for FM/non-FM/FM trilayers. At low temperatures and in multilayers with many FM/non-FM repetitions and thus a large number of interfaces the GMR ratio can exceed 200%.

11.2 Microscopic picture: Spin-dependent scattering

The mechanism leading to GMR can be understood within Mott's two-channel model (Sec. 9.3). Conduction electrons propagate due to their Fermi velocity distribution with high speed but arbitrary direction through the layered structure. A current results from a much smaller drift velocity in the direction of the applied electric field. In Fig. 30 trajectories between two reflections at outer surfaces are shown with scattering events in between. In order not to confuse the picture the changes in direction due to the scattering events are suppressed. Because of the dominance of the Fermi velocity [see schematic electron velocity distributions in Fig. 30(c)], the schematic representation and the substitutional circuit diagrams in Fig. 30 hold for both CIP and CPP geometry. We assume that minority electrons (spin antiparallel to the local magnetization) are scattered more strongly in the FM and at the FM/NM interfaces than majority electrons, and $r_{\rm min} > r_{\rm maj}$ holds for the channel resistances $r_{\rm min,maj} \propto \sigma^{-1}_{\rm min,maj}$. Thus, for parallel alignment of the magnetizations in Fig. 30(a), spin-up electrons are only weakly scattered. This leads to a low resistance for spin-up channel [green Fig. 30(a)], which short-circuits the larger resistance of the spin-down channel (black) in the circuit diagram representing the two-channel model in the lower part of Fig. 30. For antiparallel alignment of the magnetizations [Fig. 30(b)] both spin-up and spin-down electrons are majority electrons on one side of the spacer and minority

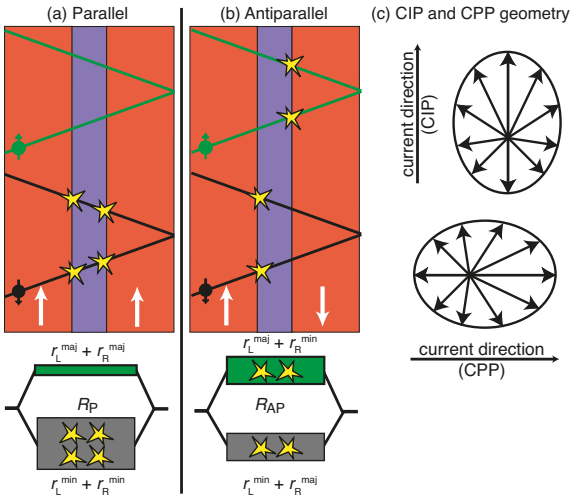


Fig. 30. *Simplified picture of the GMR effect. Only minority electrons are scattered as indicated by the stars. Majority electrons are not scattered and cause a short-circuit for parallel (a) but not antiparallel (b) magnetizations. The equivalent circuit diagrams in the lower part yield $R_P < R_{AP}$. $r_{L,R}^{maj}$ and $r_{L,R}^{min}$ denote the majority and minority equivalent resistances of the left and right side of the spacer. (c) Schematic and exaggerated total electron velocity distributions due to the Fermi velocity distribution and the superimposed drift velocity for CIP and CPP configurations, respectively.*

electrons on the other side. Hence, there are scattering events for both spin channels and the total resistance is larger than for parallel alignment. Up to now we have assumed that the diffusion trajectory of a conduction electron covers both FM layers and that the spin direction of the electron is conserved inside the spacer layer. Since there is scattering in the spacer layer, this assumption only holds when the spacer layer is thinner than the electron mean free path for the CIP geometry or the spin diffusion length for CPP geometry, for which the drift motion along the applied electric field ensures that the electron crosses the interfaces. These conditions imply that GMR only exists for spacer thicknesses of the order of a few nanometers at most.
As discussed in Sec. 9.3 the origin of the spin scattering asymmetry $r_{\text{min}} \neq r_{\text{maj}}$ can originate from the spin-dependent DOS at the Fermi level of the FM (Fig. 8) or the spin-dependent mobility. It can be expressed by a spin scattering asymmetry parameter β

$$\frac{r_{\text{min}}}{r_{\text{maj}}} = \frac{1+\beta}{1-\beta}, \tag{90}$$

where $|\beta| \leq 1$. In general, there are two contributions to the spin-dependent resistance, one arising from the bulk of the FM layers and the other from the interfaces with the spacer layer. In a trilayer system, the two magnetic layers on each side can be described by a total scattering spin asymmetry parameter (β_{L} and β_{R}; L and R stand for left and right). The resistances R_{P} and R_{AP} in Eq. (89) can be considered as a parallel connection of the two spin channels as shown at the bottom of Fig. 30. Each channel in turn is described by the sum of the contributions from the left ($r_{\text{L}}^{\text{maj,min}}$) and right ($r_{\text{R}}^{\text{maj,min}}$) half of the trilayer. Plugging the resulting expressions for R_{P} and R_{AP} into Eq. (89) and after some rearrangement, one obtains

$$\frac{\Delta R}{R_{\text{P}}} = C\beta_{\text{L}}\beta_{\text{R}}, \tag{91}$$

where C is a constant, which is always positive. Therefore, the product $\beta_{\text{L}}\beta_{\text{R}}$ determines whether the GMR effect is normal ($\beta_{\text{L}}\beta_{\text{R}} > 0$) or inverse ($\beta_{\text{L}}\beta_{\text{R}} < 0$). Obviously, for a symmetric trilayer, $\beta_{\text{L}} = \beta_{\text{R}}$ holds, and the GMR is always normal. Numerous experiments showed that

for the $3d$ transition metals and their alloys, the sign of the β's can be obtained from the slopes of the Slater-Pauling curve in Fig. 12. The negative slope on the right-hand side, where Co is located, signifies that adding more electrons decreases the magnetic moment per atom. This requires that the minority DOS at the Fermi level is larger than the majority DOS, $N^{\text{min}} > N^{\text{maj}}$ [see Fig. 11(b) for the example of Co]. These different densities of the final states for scattering events yield $r_{\text{min}} > r_{\text{maj}}$ or $\beta > 0$. Correspondingly, the positive slope in the left part of Fig. 12 predicts $\beta < 0$. This rule holds for the bulk scattering spin asymmetries in the alloys of the metals A and B with a composition given by the average number of electrons per atom (abscissa in Fig. 12) but also for the interface scattering spin asymmetries of A/B interfaces. For instance, CoCr alloys as well as Co/Cr interfaces have a negative β. This relation between the Slater-Pauling curve and the signs of the GMR effect is observed in CIP and CPP geometry and confirms that spin-dependent scattering due to the exchange-split DOS of the FM layers is the predominant mechanism for GMR.

12 Tunneling Magnetoresistance (TMR)

The first tunnel magnetoresistance (TMR) effect of 14% was observed already in 1975 by Jullière [29] in Fe/Ge/Co junctions. However, the effect was only observable at liquid He temperature. Triggered by the success of GMR, FM/insulator/FM structures were revisited in 1995 and up to 18% TMR effect could be observed for the first time at room temperature [30, 31]. By now the record value for AlO_x barriers of 70% at room temperature is achieved by using amorphous CoFeB electrodes. However, much higher TMR values have been found in epitaxial structures with MgO(001) barriers and will be discussed in Sec. 12.3

12.1 Phenomenological description

The basic configuration for tunnel magnetoresistance (TMR) consists of two FM electrodes, usually realized as thin films, separated by an insulating barrier as shown in the upper part of Fig. 31. If a voltage V (several tens to hundreds mV) is applied across the stack, a small quantum-mechanical tunneling current can flow across the barrier. This means that -unlike GMR- the TMR effect is observed only in CPP geometry. The magnitude of the tunneling current is related to the overlap of the exponentially decaying wave functions inside the barrier. Therefore, the current exponentially decreases with the barrier thickness. Typical barrier thicknesses are of the order of 0.5 to 1.5 nm. Electron tunneling is discussed in detail in lecture A9 by Daniel Wortmann.
The tunneling resistance is found to depend on the relative orientation of the magnetizations on both sides of the barrier. The magnitude of the TMR effect is determined in the same way as for GMR [compare Eq. (89)]

$$\frac{\Delta R}{R_{\text{P}}} = \frac{R_{\text{P}} - R_{\text{AP}}}{R_{\text{P}}}. \tag{92}$$

12.2 Microscopic picture: Spin-dependent tunneling

The TMR effect can be understood on the basis of **spin-polarized tunneling**. If the spin is conserved during tunneling, a spin-up (spin-down) electron can only tunnel from an initial spin-up (spin-down) state to an unoccupied spin-up (spin-down) final state. TMR arises from the imbalance between the number of spin-up and spin-down electrons that contribute to the

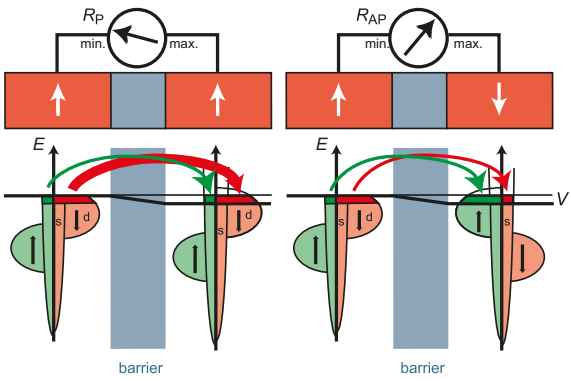


Fig. 31. *Assuming energy and spin conservation during the tunneling process, the tunneling current can be decomposed into spin-up (green arrows) and spin-down (red arrows) contributions. Their magnitudes (thickness of the arrows) are determined by the number of available initial and final states for each spin channel, here given by the simplified DOS of 3d transition metals. Hence, the resistance of the tunnel junction depends on the alignment of the magnetizations.*

tunneling current. Therefore, we define the spin polarizations P_{L} and P_{R} of the left and right electrodes

$$P_{\mathrm{L,R}} = \frac{N^{\uparrow}_{\mathrm{L,R}} - N^{\downarrow}_{\mathrm{L,R}}}{N^{\uparrow}_{\mathrm{L,R}} + N^{\downarrow}_{\mathrm{L,R}}}, \tag{93}$$

where $N^{\uparrow}_{\mathrm{L,R}}$ and $N^{\downarrow}_{\mathrm{L,R}}$ denote the number of states in an energy window at the Fermi level with a width given by the applied voltage V. Only states within this window (dark green and red colored areas in Fig. 31) can contribute to the tunneling current. In Fig. 31 a positive voltage V is applied to the right electrode. The green and red bent arrows represent the spin-up and spin-down tunneling currents, respectively, with their thickness indicating the magnitude of the currents. For instance, the spin-up current in the parallel configuration is proportional to the product $N^{\uparrow}_{\mathrm{L}} N^{\uparrow}_{\mathrm{R}}$ (green arrow in the left hand part). Obviously, the parallel alignment on the left hand side of Fig. 31 gives rise to a larger total current and, thus, to the smaller tunneling resistance. R_{P} and R_{AP} are inversely proportional to the total current (*i.e.* the sum of the spin-up and spin-down currents) and can be written as

$$R_{\mathrm{P}} \propto \frac{V}{N^{\uparrow}_{\mathrm{L}} N^{\uparrow}_{\mathrm{R}} + N^{\downarrow}_{\mathrm{L}} N^{\downarrow}_{\mathrm{R}}} \quad ; \quad R_{\mathrm{AP}} \propto \frac{V}{N^{\uparrow}_{\mathrm{L}} N^{\downarrow}_{\mathrm{R}} + N^{\downarrow}_{\mathrm{L}} N^{\uparrow}_{\mathrm{R}}}. \tag{94}$$

Inserting these expressions into Eq. (92) and some rearranging yields $\frac{\Delta R}{R_{\mathrm{P}}}$ as a function of the polarizations defined in Eq. (93)

$$\frac{\Delta R}{R_{\mathrm{P}}} = \frac{R_{\mathrm{P}} - R_{\mathrm{AP}}}{R_{\mathrm{P}}} = \frac{2 P_{\mathrm{L}} P_{\mathrm{R}}}{1 - P_{\mathrm{L}} P_{\mathrm{R}}}. \tag{95}$$

This expression for the TMR ratio is called **Jullière's TMR formula** after the inventor of this model [29]. If ΔR is positive (negative), the TMR effect is called normal (inverse). In Fig. 31 the effect turns out to be normal, but an inverse effect can result if the magnetic electrodes on both sides of the barrier were different in such a way that the P_{L} and P_{R} have opposite signs [see Eq. (95)]. Examples will be given below. The TMR effect usually decreases as a function of bias voltage and temperature. Spin scattering in the barrier, DOS effects as well as the excitation of spin waves are possible reasons.

12.3 Beyond Jullière's model

Obviously, Jullière's model is very simple, and it is not surprising that several experimental observations cannot be consistently explained in the framework of this model. For instance, it is not clear how to obtain the relevant values for P_{L} and P_{R}. Polarizations determined from TMR measurements using the Jullière relation [Eq. (95)] are sometimes in strong disagreement (in some cases even concerning the sign) with polarizations determined by other techniques. For TMR, only the polarization of the electronic states right at the interfaces (including possible interface states) and in an energy window at the Fermi level with a width given by the applied voltage V are important, and the tunneling current is dominated by states with maximum momentum perpendicular to the barrier. Furthermore, the Jullière model does not consider realistic barriers with band structures different from the vacuum. Note that the barrier properties do not explicitly appear in Eq. (95). An impressive demonstration of the influence of the barrier material has been given in Ref. [32] and is shown in Fig. 32. In these experiments the two FM electrodes are Co and $La_{0.7}Sr_{0.3}MnO_3$, but different barrier materials are used. For $SrTiO_3$ and $Ce_{0.69}La_{0.31}O_{1.845}$ barriers the TMR was found to be inverse [Figs. 32(a) and (b)], whereas it was normal for Al_2O_3 and $Al_2O_3/SrTiO_3$ barriers [Figs. 32(c) and (d)] clearly proving that TMR strongly depends on the barrier material. The spin polarizations P_{L} and P_{R} must thus be related to interface states, which play a major role for the chemical bonding at the interfaces.
The limited validity of Jullière's model also became obvious in 2004 when extremely high TMR ratios of up to 220% at room temperature were reported for epitaxial [34] or highly oriented [35] MgO(001) barriers and Fe or CoFe electrodes. The high TMR ratios translate with the Jullière TMR formula [Eq. (95)] into spin polarizations $P_{\mathrm{L,R}} \approx 70\%$. These values are definitely too high to be identified with bulk spin polarizations of Fe or CoFe alloys, which typically are about 40%. The experiments, however, confirm a theoretical prediction [33] that single-crystalline,

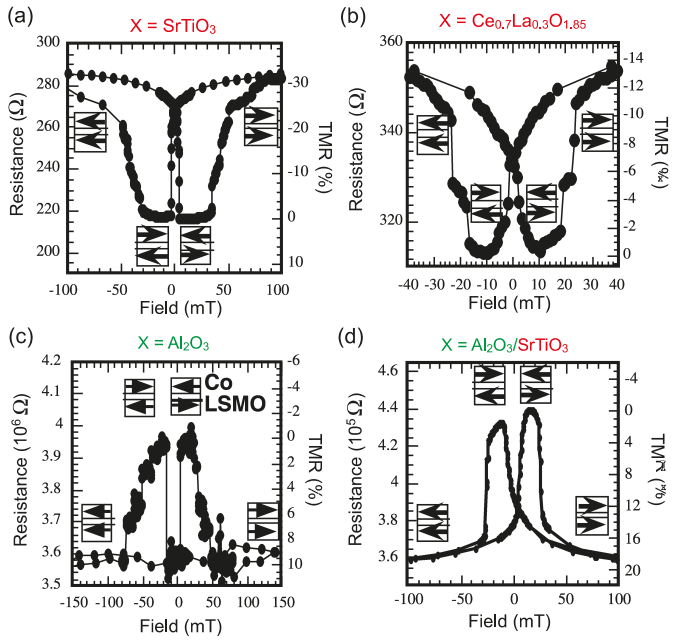


Fig. 32. *TMR of $Co/X/La_{0.7}Sr_{0.3}MnO_3$ structures with X = (a) $SrTiO_3$, (b) $Ce_{0.69}La_{0.31}O_{1.845}$, (c) Al_2O_3, and (d) = $Al_2O_3/SrTiO_3$. The material-dependent change from the inverse (a,b) to the normal (c,d) TMR effect indicates the influence of the Co/insulator interface. After Ref. [32].*

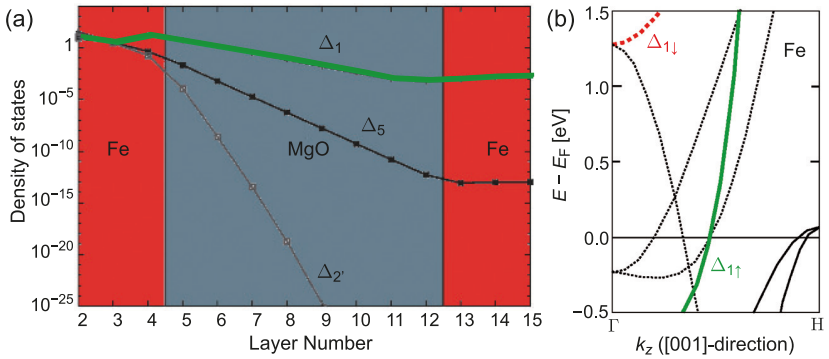


Fig. 33: *(a) States with Δ_1-symmetry decay slowest in the MgO barrier. The DOS of all other symmetries is suppressed by orders of magnitude. (b) In Fe the bands with Δ_1-symmetry are exchange-split and only the $\Delta_1^{\uparrow}$-band crosses the Fermi level. After Ref. [33].*

epitaxial MgO barriers in combination with Fe(001) electrodes would yield huge TMR ratios of hundreds of percent. In contrast to amorphous AlO_x barriers the tunneling across epitaxial MgO is coherent, meaning that the symmetry of the states is conserved. Therefore, specific features in the band structures of MgO and Fe can be exploited: (i) The complex band structure of MgO (see lecture A9 by Daniel Wortmann) yields a much smaller exponential decay in the barrier for the electronic states with Δ_1 symmetry compared to all other symmetries [Fig. 33(a)]. Therefore, the tunneling current in Fe/MgO/Fe(001) is predominantly carried by Δ_1-states. (ii) In Fe, the $\Delta_1^{\uparrow}$ and $\Delta_1^{\downarrow}$ bands are strongly exchange-split, and only the majority $\Delta_1^{\uparrow}$ band crosses the Fermi level [Fig. 33(b)]. This leads to a strong spin selection in the tunneling process, and hence a TMR ratio as high as 600% at 300 K (1144% at 5 K) [36]. Although the combination of MgO with Fe-based alloys with bcc structure (*e.g.* CoFe or CoFeB) seems to be a particular case this material system is by now applied in magnetic random access memories (MRAM) and read-heads of hard-disk drives.

13 Spin-transfer torque (STT)

The magnetoresistance effects GMR and TMR describe the influence of the relative magnetization alignment between adjacent magnetic layers on the current flow, *i.e.* the resistance. The **spin-transfer torque (STT)** effects to be discussed in this section represent a reciprocal interaction (Fig. 34): In a device with inhomogeneous magnetization profile a strong current can transfer spin momentum between different parts of the device and thus exerts a torque on the local magnetization and thereby influences the magnetization configuration. Current-induced magnetization switching, excitation of steady magnetization oscillations, and current-induced domain wall motion are experimentally observable manifestations of the **current-driven magnetization dynamics** due to STT.

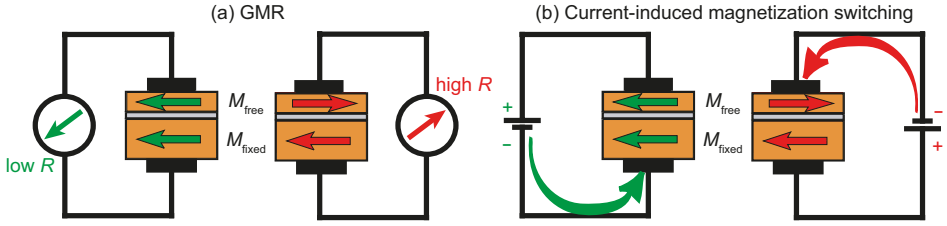


Fig. 34: *Phenomenology of (a) GMR and (b) current-induced magnetization switching as a manifestation of STT effects. (a) The electric resistance of a trilayer structure consisting of two FM separated by a non-magnetic, metallic interlayer depends on the alignment of the layer magnetizations. (b) The stable alignment of the magnetizations depends on the polarity of the current flowing perpendicularly through the trilayer.*

13.1 Phenomenological description of STT

In 1996 Slonczewski [38] and Berger [39] predicted that a spin-polarized current propagating into a FM layer exerts a torque on the layers' magnetization, due to the exchange interaction between the electrons and the local magnetic moments. In layered metallic systems with alternating magnetic and non-magnetic layers, a current flowing perpendicular to the plane of the layers (CPP-geometry) is polarized by one FM layer and transfers spin angular momentum to another FM layer, where the transferred momentum acts as a torque on the magnetization. This effect is called **spin-transfer torque**. For this torque to be sufficient to perturb the magnetization from equilibrium, large current densities ($> 10^7$ A/cm^2) are required. If two stable equilibria for the magnetization exist (*e.g.* due to a uniaxial anisotropy), the STT can reversibly switch the magnetization between the two equilibrium positions. This magnetic switching scheme does not require an external magnetic field. Its phenomenology is shown in Fig. 34(b). We consider two FM layers separated by a non-FM spacer with a thickness less than its spin diffusion length. The FM layers are different in such a way (*e.g.* thickness or coercive field), that one of them can

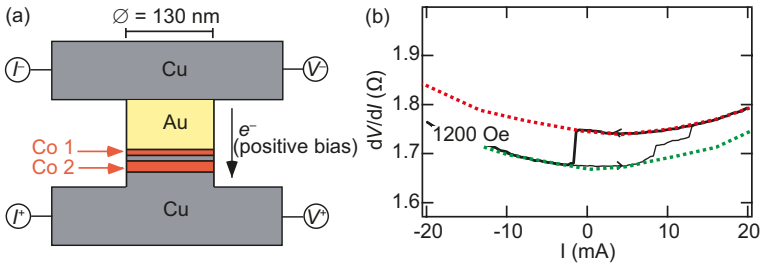


Fig. 35: *(a) Schematic pillar device with two Co layers (Co 1 and Co 2) separated by a 6 nm thick Cu layer. (b) The dV/dI measurements as a function of the current through the column device yields the relative alignment of the magnetic layers via the GMR effect. At positive bias electrons flow from Co 1 to Co 2 layer. For large enough current Co 1 switches to antiparallel alignment as indicated by the higher resistance (red line). For negative bias parallel alignment and a lower resistance (green line) is observed. An external field of 1200 Oe is applied to fix the magnetization direction of Co 2. After Ref. [37].*

be remagnetized more easily than the other. We distinguish the two layers in the following by calling them *free* and *fixed* and draw them as a thinner and thicker layer, respectively. When electrons flow from the fixed to the free layer, the magnetization of the free layers aligns parallel to the magnetization of the fixed layer and this alignment is stabilized. When the current direction is reversed, however, the antiparallel alignment is more stable and adopted. Thus, a magnetization reversal can be induced by reversing the polarity of the DC current flowing through the layers.
An experimental arrangement for the observation of current-induced switching is displayed in Fig. 35(a). It consists of a column of layers of various materials stacked on top of each other. A current can be fed in by leads I^- and I^+, and the voltage drop is measured at V^- and V^+. There is a thin Co layer (Co 1) with a thickness of 2.5 nm and a thick Co layer (Co 2) of 10 nm thickness. The Cu spacer in between is 6 nm thick. The lateral diameter of the column is only about 100 nm to reach the necessary high current density of 10^8 A/cm^2. As shown in Fig. 35(b), the relative orientation of the Co layers can be measured via the GMR effect (Sec. 11) of the Co 1/Cu/Co 2 trilayer. At negative bias electrons flow from Co 2 to Co 1 layer and stabilize the parallel magnetization alignment, which yields a low dV/dI. At positive bias the parallel alignment is destabilized, Co 1 switches to the antiparallel alignment at a sufficiently large current, and dV/dI increases. Upon reducing the current [thick line in Fig. 35(b)], hysteretic behavior is observed such that Co 1 switches back at a smaller current. The magnetization direction of the thicker Co 2 layer is fixed by an external field.

13.2 Physical picture of STT: Absorption of the transverse spin current component

Being aware of the high current densities, one might suppose that the Oersted field generated by the current is responsible for the switching behavior. The circumferential Oersted field favors for a magnetic disk the so-called magnetic vortex state, which is characterized by an in-plane circulation of the magnetization and a small perpendicularly magnetized core region. This arrangement minimizes the stray field energy because no magnetic flux penetrates the surface of the disk except in the small core region. However, switching into vortex states due to the Oersted field has the wrong symmetry: Both current polarities would lead to a vortex-like magnetization state but with opposite sense of rotation. Nevertheless, they would result in the same GMR response and a symmetric behavior for positive and negative currents is expected in clear contrast to the data in Fig. 35(b). Furthermore, the strongest Oersted field occurs at the pillar circumference and scales like I/d, where I is the current and d the pillar diameter. The STT effect scales like the current density I/d^2 and therefore becomes stronger below a certain structure size d_c. Theoretical estimates and experiments suggest a d_c of the order of 100 nm. This fundamental size restriction coincides with the possibilities of e-beam lithography and at the same time yields the needed current densities at technically convenient current amplitudes of the order of mA. In practice one always has to be aware of the presence of the Oersted field and has to take its possible influence into account.
In order to develop a physical picture for STT, we first consider a spin-polarized current that enters a FM from a metallic non-magnet [Fig. 36(a)] and is polarized along an axis tilted by the angle θ with respect to the magnetization $\vec{M}$ of the FM. The (normalized) wave function of a polarized electron can be written as a superposition of spin-up and spin-down spinor components with respect to the quantization axis defined by $\vec{M}$. The amplitudes are $\cos(\theta/2)$ and $\sin(\theta/2)$, respectively, and correspond to a transverse component of the spin vector given by $\sin(\theta)$. At

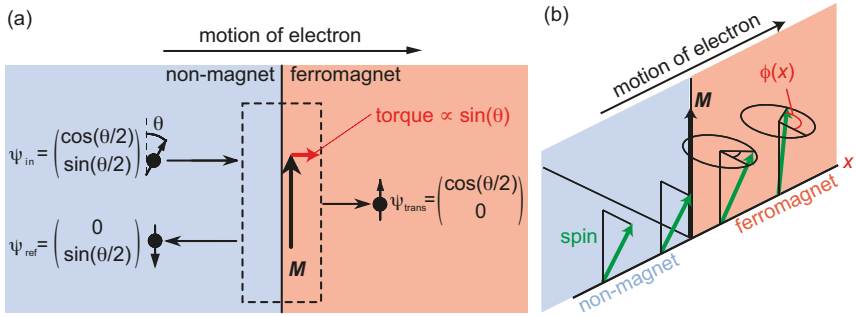


Fig. 36: *Two effects contributing to the absorption of the transversal spin current component at the interface between a non-magnet and a ferromagnet. (a) Spin filtering: The incoming Ψ_{in}, transmitted Ψ_{trans}, and reflected Ψ_{ref} spinors for the idealized case of perfect spin filtering are indicated. The absorbed transversal spin current is proportional to $\sin\theta$ and acts as a torque on the interface magnetization. (b) Spatial precession of the spin in the ferromagnet: The phase ϕ is constant in the non-magnet, but increases in the ferromagnet with distance x from the interface.*

the interface to the FM the potential experienced by the electron changes and becomes spin-dependent. Therefore, the transmitted and reflected wave functions are different superpositions of spin-up and spin-down spinor components compared to the incident wave function. This leads unavoidably to different transverse spin components and thus to a discontinuity in the transverse spin current. The missing transverse spin current is absorbed at the interface and acts as a torque on the magnetization. This effect occurs for each electron individually and is called **spin filtering** [38]. Figure 36(a) shows the spinors in the extreme case of perfect spin filtering. In realistic cases, roughly 50% of the transversal component is absorbed, and the transmitted as well as reflected currents still carry transversal components [40]. The actual current polarization of the transmitted and reflected currents is obtained by summing over all conduction electrons. This introduces two additional effects. The first arises because the reflection and transmission amplitudes at the interface are complex and k-dependent. This means that the spin of an incoming electron rotates upon reflection and transmission by a k-dependent angle. The cancellation, which occurs when we sum over all k-states, reduces the net outgoing transverse spin current. This is an entirely quantum-mechanical phenomenon, for which there is no classical analog. A second effect arises because spin-up and spin-down electrons on the Fermi surface have the same wave vector $k_\uparrow = k_\downarrow$ in the non-magnet but no longer when they enter the FM, $\Delta k = k_\uparrow - k_\downarrow \neq 0$. This is a consequence of the spin-split DOS. The two components are coherent, and a spatial phase $\phi(x) = \phi_0 + \Delta k \cdot x$ builds up [Fig. 36(b)] corresponding to a precession of the spin vector in space. The precession frequency is k-dependent, *i.e.* varies with the position of the considered state on the Fermi surface. Therefore, when we sum over all conduction electrons, the transverse spin component is almost completely canceled out after propagation by a few lattice constants into the FM. Taking all three effects – (i) spin filtering, (ii) rotation of the reflected and transmitted spin, and (iii) spatial precession of the spin in the FM – together, to a good approximation, the transverse component of the transmitted and reflected spin currents are zero for most systems of interest. Thus, the incoming transverse spin current is absorbed by the interface and acts as a current-induced torque on the magnetization.

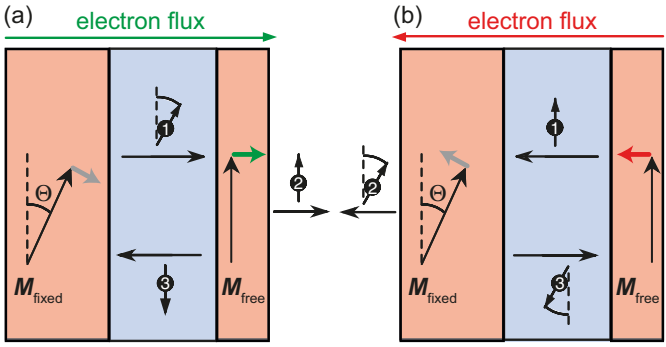


Fig. 37: *Orange regions represent the two FM layers. Due to the assumed asymmetry $\vec{M}_{fixed}$ does not respond to the torque (short gray arrows) acting on it, whereas $\vec{M}_{free}$ can follow the torque (short green and red arrows). The numbers in the spins refer to the sequence of the description. (a) and (b) show the situation for opposite electron flux directions, which result in a stabilization or destabilization, respectively, of the parallel alignment.*

A comprehensive theoretical treatment of these effects is given in Ref. [40].

Up to now we have assumed that the incident current is polarized. In the experiment this can be achieved by a second ferromagnetic layer with a slightly tilted magnetization (angle θ). The spin polarization is not modified in the non-magnetic spacer layer provided the spacer layer thickness is below its spin diffusion length to prevent significant depolarization by spin-flip scattering. In Fig. 37 we consider a trilayer structure very similar to the experimental setup of Fig. 35(a). In Fig. 37(a) the electrons flow from the fixed to the free layer. A current polarized by the fixed layer (1) hits the free layer and transfers its transversal component as a torque to the free layer. Part of the current is transmitted (2) and another part is reflected (3). This reflected current can now be considered as a polarized current impinging on the fixed layer. Again, the transversal component will be absorbed and acts as a torque on the fixed layer. However, due to the assumed asymmetry the fixed layer will resist to the torque, and only $\vec{M}_{\text{free}}$ starts to rotate in order to reach the stable parallel alignment with $\vec{M}_{\text{fixed}}$. For the opposite direction of the electron flux in Fig. 37(b) we obtain a similar situation but the torques point in opposite directions. Therefore, the stable state corresponds to the antiparallel alignment of $\vec{M}_{\text{free}}$ and $\vec{M}_{\text{fixed}}$. Note, that in this case the torque on $\vec{M}_{\text{free}}$ arises from the current, which first has been reflected from the fixed layer. Obviously, the asymmetry (fixed *versus* free) plays an important role, which is very reasonable because left and right cannot be distinguished for the symmetric case.

This consideration is valid independent of what gives rise to the spin-polarized current. The above picture holds for ballistic transport due to an applied bias voltage. In the case of diffusive transport there is also a contribution due the spin accumulation that builds up at the FM/NM interfaces as discussed in Sec. 9.3. The gradient of the spin accumulation gives rise to a spin current with a polarization that in general is not collinear with the magnetization in the FM, and the processes discussed above apply: The transversal component of the diffusive spin current is also absorbed and acts on the magnetization like a torque [25].

13.3 Extended Gilbert equation and spin-torque oscillators

In order to address the question how this torque influences the dynamics of the macroscopically observable magnetization, we have to consider the **Gilbert equation**,

$$\frac{d\vec{m}}{dt} = \underbrace{-\gamma\,\vec{m}\times\vec{H}_{\text{eff}}}_{\propto\frac{d\vec{M}_P}{dt}} + \underbrace{\alpha\,\vec{m}\times\frac{d\vec{m}}{dt}}_{\propto\frac{d\vec{M}_D}{dt}}, \tag{96}$$

which is the equation of motion for a magnetization $\vec{m}$ in an effective field $\vec{H}_{\text{eff}}$. Here, $\vec{m} = \vec{M}/M_{\text{S}}$ is the reduced magnetization, γ the gyromagnetic ratio, and α the phenomenological Gilbert damping constant. The effective field is the negative variational derivative of the total areal energy density E_{tot} comprising contributions from exchange, anisotropy, stray field, and Zeeman energy with respect to the magnetization, $\vec{H}_{\text{eff}} = -\frac{1}{\mu_0}\frac{\delta E_{\text{tot}}}{\delta \vec{M}}$. The first term in Eq. (96) describes the precessional motion of $\vec{m}$ about H_{eff} and the second term the damping, which forces $\vec{m}$ to relax to the lowest energy configuration, $\vec{m}||H_{\text{eff}}$ (gray arrows in Fig. 38).

Slonczewski [38] expressed the current-induced STT $d\vec{M}_{\text{free}}/dt$ acting on the free layer as

$$\frac{1}{M_S}\frac{d\vec{M}_{\text{free}}}{dt} = \frac{d\vec{m}_{\text{free}}}{dt} = \frac{I}{A}\cdot g(\theta)\cdot\vec{m}_{\text{free}}\times(\vec{m}_{\text{free}}\times\vec{m}_{\text{fixed}}), \tag{97}$$

where I/A is the current density, $g(\theta)$ is the material-dependent STT efficiency function, which is a measure for the conversion of current into STT. In general, it depends on the angle θ between $\vec{M}_{\text{free}}$ and $\vec{M}_{\text{fixed}}$. The materials enter *via* the spin polarization P, volume and interface resistances, and other transport properties. The double cross product is indeed proportional to $\sin\theta$ and, thus, the absorbed transversal component of the spin current as discussed in the context of Fig. 36(a). The linear dependence on I yields the reversed torque upon reversing the current direction. The direction of $d\vec{M}_{\text{free}}/dt$ for the two current polarities is shown in Fig. 38 by the red and green arrows, respectively. The latter case is more interesting, because the conventional damping torque $d\vec{M}_{\text{D}}/dt$ may be compensated or even overcome by the STT term $d\vec{M}_{\text{STT}}/dt$. In this case the precession amplitude increases and $\vec{M}_{\text{free}}$ is destabilized, which leads to magnetization switching or the excitation of steady-state oscillatory modes.

In the phenomenological description of current-induced magnetization switching, we have considered the STT and damping terms of the Gilbert equation, but neglected the precessional term. A more complete analysis taking all terms into account shows that the switching process after applying a DC current of the correct polarity starts with the excitation of a precessional motion about the initial state. The cone angle of the trajectory increases steadily under the action of the STT, which opposes the restoring Gilbert torque. When $\vec{M}$ reaches the position, where a

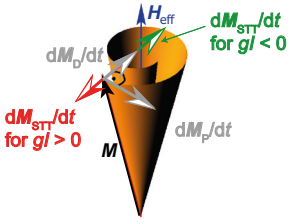


Fig. 38. *Motion of a magnetization vector $\vec{M}$ in an effective field $\vec{H}_{eff}$. The first term in Eq. (96) gives rise to the tangential torque $d\vec{M}_P/dt$ driving the precession and the second term $d\vec{M}_D/dt$ causes the damping. The spin-transfer torque $d\vec{M}_{STT}/dt$ can point along the Gilbert damping $d\vec{M}_D/dt$ (green) or opposite to it (red). In the latter case it can destabilize $\vec{M}$ and induce magnetization switching or microwave oscillations.*

potential maximum separates the initial and the final states, switching occurs and $\vec{M}$ relaxes towards the final state, now on a precessional trajectory with decreasing cone angle. This process only happens if the external field is lower than the coercive field of the free layer. The shape or magnetocrystalline anisotropy then gives rise to at least two stable states, and the current-induced STT can cause switching between them. If the external field exceeds the coercivity only one stable magnetization state exists, namely parallel to the external field, and switching is not possible for either current polarity. For one polarity the system is not excited at all, whereas for the other polarity it enters a **steady-state oscillatory motion**, which is characterized by the equilibrium between $d\vec{M}_{\mathrm{D}}/dt$ and $d\vec{M}_{\mathrm{STT}}/dt$. In this way, $\vec{M}_{\mathrm{free}}$ can be driven into new types of oscillatory dynamic modes, which are not attainable with magnetic fields alone. Any oscillatory motion of the free layer with respect to the fixed layer results, due to the GMR or TMR (Secs. 11 and 12) effect, in a variation of the resistance. Therefore, the DC current generates a oscillatory voltage signal with typical frequency in the GHz range that can be measured with a HF spectrum analyzer. Nanomagnets driven by spin-polarized currents have the potential to serve as nanoscale, on-chip microwave sources or oscillators, tunable by field and current over a wide frequency range. These devices are called **spin-torque oscillators**.

13.4 Current-driven domain wall motion

Up to now we have assumed that the magnetic element, which is subject to the STT, displays a uniform magnetization pattern in the static state. However, STT also occurs when a current passes through a non-uniformly magnetized object. A domain wall (Sec. 8) is by definition a non-uniform magnetization pattern. A thin and narrow magnetic wire is divided by domain walls into sections of opposite magnetization directions. A current flowing along the wire has to repolarize each time after passing a domain wall in order to adjust to the local magnetization. This situation is very similar to Fig. 37, except that the domain wall plays the role of the spacer layer. The total action of the current-induced torques on the magnetization leads to a motion of a domain wall in the direction of the electron flux. **Current-driven domain wall motion** is employed in a novel magnetic storage concept called **Magnetic Racetrack Memory** to move magnetic domain walls, which carry the stored information, along a magnetic wire.

Acknowledgment

During the preparation of this lecture I made use of inspiring sources in Refs. [41, 42, 43]. I also thank Paul Bechthold and Frank Matthes for proofreading this manuscript.

References

[1] W. Gilbert, *de Magnete (Translated by P. F. Mottelay)* (Dover Publications Inc., 1958).

[2] J. C. Maxwell, Phil. Trans. Roy. Soc. London **155**, 459 (1865).

[3] P. Grünberg, R. Schreiber, Y. Pang, M. B. Brodsky, and H. Sowers, Phys. Rev. Lett. **57**(19), 2442 (1986).

[4] M. N. Baibich, J. M. Broto, A. Fert, F. N. V. Dau, F. Petroff, P. Etienne, G. Creuzet, A. Friedrich, and J. Chazelas, Phys. Rev. Lett. **61**(21), 2472 (1988).

[5] G. Binasch, P. Grünberg, F. Saurenbach, and W. Zinn, Phys. Rev. B **39**(7), 4828 (1989).

[6] I. Dzyaloshinskii, J. Phys. Chem Solids **4**, 241 (1958).

[7] T. Moriya, Phys. Rev. **120**, 91 (1960).

[8] M. Bode, M. Heide, K. von Bergmann, P. Ferriani, S. Heinze, G. Bihlmayer, A. Kubetzka, O. Pietzsch, S. Blügel, and R. Wiesendanger, Nature **447**, 190 (2007).

[9] S. Heinze, K. von Bergmann, M. Menzel, J. Brede, A. Kubetzka, R. Wiesendanger, G. Bihlmayer, and S. Blügel, Nature Phys. **7**, 713 (2011).

[10] U. K. Rößler, A. N. Bogdanov, and C. Pfleiderer, Nature **442**, 797 (2006).

[11] G. Chen, J. Zhu, A. Quesada, J. Li, A. T. N'Diaye, Y. Huo, T. P. Ma, Y. Chen, H. Y. Kwon, C. Won, Z. Q. Qiu, A. K. Schmid, *et al.*, Phys. Rev. Lett. **110**, 177204 (2013).

[12] S.-W. Cheong and M. Mostovoy, Nature Mater. **6**, 13 (2007).

[13] J. F. Janak, Phys. Rev. B **16**, 255 (1977).

[14] E. Y. Tsymbal and D. G. Pettifor, *Solid State Physics: Advances in Research and Applications*, vol. 56 (Academic Press, 2001).

[15] V. L. Moruzzi, J. F. Janak, and A. R. Williams, *Calculated Electronic Properties of Metals* (Pergamon Press, New York, 1978).

[16] A. Barthélémy, A. Fert, and F. Petroff, *Handbook of Magnetic Materials* (Elsevier, Amsterdam, 1999), vol. 12, chap. Giant Magnetoresistance in Magnetic Multilayers.

[17] S. LaShell, B. A. McDougall, and E. Jensen, Phys. Rev. Lett. **77**, 3419 (1996).

[18] M. Hoesch, M. Muntwiler, V. N. Petrov, M. Hengsberger, L. Patthey, M. Shi, M. Falub, T. Greber, and J. Osterwalder, Phys. Rev. B **69**, 241401 (2004).

[19] S. Datta and B. Das, Appl. Phys. Lett. **58**(7), 665 (1990).

[20] F. J. A. den Broeder, W. Hoving, and P. J. H. Bloemen, J. Magn. Magn. Mater. **93**, 562 (1991).

[21] D. Mauri, H. C. Siegmann, P. S. Bagus, and E. Kay, J. Appl. Phys. **62**(7), 3047 (1987).

[22] A. P. Malozemoff, Phys. Rev. B **35**(7), 3679 (1987).

[23] N. C. Koon, Phys. Rev. Lett. **78**(25), 4865 (1997).

[24] N. F. Mott, Proc. Roy. Soc. **153**, 699 (1936).

[25] M. D. Stiles and A. Zangwill, J. Appl. Phys. **91**(10), 6812 (2002).

[26] J. Smit, Physica **16**(6), 612 (1951).

[27] I. A. Campbell, A. Fert, and O. Jaoul, J. Phys. C: Metal Phys. Suppl. **1**, S95 (1970).

[28] B. Dieny, J. Magn. Magn. Mater. **136**, 335 (1994).

[29] M. Julliere, Phys. Lett. **54A**, 225 (1975).

[30] T. Miyazaki and N. Tezuka, J. Magn. Magn. Mater. **139**, L231 (1995).

[31] J. S. Moodera, L. R. Kinder, T. M. Wong, and R. Meservey, Phys. Rev. Lett. **74**(16), 3273 (1995).

[32] J. M. De Teresa, A. Barthélémy, A. Fert, J. P. Contour, F. Montaigne, and P. Seneor, Science **286**, 507 (1999).

[33] W. H. Butler, X.-G. Zhang, T. C. Schulthess, and J. M. MacLaren, Phys. Rev. B **63**, 054416 (2001).

[34] S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, and K. Ando, Nature Mater. **3**, 868 (2004).

[35] S. S. P. Parkin, C. Kaiser, A. Panchula, P. M. Rice, B. Hughes, M. Samant, and S.-H. Yang, Nature Mater. **3**, 862 (2004).

[36] S. Ikeda, J. Hayakawa, Y. Ashizawa, Y. M. Lee, K. Miura, H. Hasegawa, M. Tsunoda, F. Matsukura, and H. Ohno, Appl. Phys. Lett. **93**, 082508 (2008).

[37] J. A. Katine, F. J. Albert, R. A. Buhrman, E. B. Myers, and D. C. Ralph, Phys. Rev. Lett. **84**(14), 3149 (2000).

[38] J. C. Slonczewski, J. Magn. Magn. Mater. **159**, L1 (1996).

[39] L. Berger, Phys. Rev. B **54**(13), 9353 (1996).

[40] M. D. Stiles and A. Zangwill, Phys. Rev. B **66**, 014407 (2002).

[41] R. Gross and A. Marx, Lectures Notes on Solid-State Physics (Walther-Meissner-Institut 2009) and Spintronics (Walther-Meissner-Institut 2004).

[42] M. Getzlaff, *Fundamentals of Magnetism* (Springer, Heildelberg, 2008).

[43] Lecture notes of P. Bechthold, G. Bihlmayer, S. Blügel, C. M. Schneider, K. Schröder, D. Wortmann, and R. Zeller of previous IFF Spring Schools (Forschungszentrum Jülich 1998, 2005, 2009).

A 9 Electron Tunneling [1]

Daniel Wortmann, Phivos Mavropoulos
Peter Grünberg Institut and
Institute for Advanced Simulation
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Single Particle View on Quantum Transport **2**

3 Tunneling through a rectangular Barrier **4**
3.1 Analytical solution for the rectangular barrier model 4
3.2 Resonant tunneling through virtual bound states 5
3.3 Resonant tunneling through surface states 7

4 Landauer formula **8**
4.1 Interpretation of the Landauer formula 10

5 The Bardeen Approach to Tunneling **11**
5.1 Landauer conductance versus Bardeen's tunneling 14
5.2 Cu-vacuum-Cu tunneling 16

6 Complex band structure **18**

7 Applications **19**
7.1 Tunneling magneto-resistance 19
7.2 Tunneling electro-resistance 22

Lecture Notes of the 47$^{\text{th}}$ IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

The tunnel effect is ubiquitous in quantum physics. Some of its first applications are found in the theory of electron emission from metal surfaces [1] and in the modeling of the α-decay of heavy nuclei [2, 3]. In its original formulation, the tunnel effect comprises the non-vanishing probability of a particle to penetrate and traverse a "forbidden" spatial region, or a "barrier," where the potential is higher than the particle energy. In Solid State Physics, however, the energy-band theory of the elecronic structure has given a wider sense to the term of tunneling. Here, the barrier may comprise the band gap of an insulator separating two metallic leads, irrespective of the actual value of the electron potential. Still, the essence of the tunnel effect is the same, namely that the probability T to traverse the barrier (the tunneling, or transmission, probability) is in most cases controlled by two parameters: the decay parameter κ, which characterises the barrier type, and the barrier thickness d. The order of magnitude of the transmission probability is $T \sim \exp(-\kappa d)$, and this exponential control suggests the tunnel effect as an extremely sensitive probing tool. To name only a few applications, the Scanning Tunneling Microscopy and Spectroscopy (STM/STS), the Field Emission Spectroscopy of metal surfaces, or the Field Effect Transistor, the Tunneling Magneto-Resistance (TMR) and Tunneling Electro-Resistance (TER) contribute on a standard basis either to basic research or to technology.
The few aforementioned examples share some common elements: the decay parameter κ is controlled during operation through an external field[2] (electric or magnetic, depending on the application); in addition, the electron tunneling is triggered by a bias voltage between two metallic regions. Thus, the applications fall in the category of electron transport with exponentially sensitive control of the resistance. In this context, it is informative to approach the tunneling problem through electron-transport theory. Both subjects (electron transport and tunneling) are of course vast, thus we must restrict the present manuscript to specific aspects. Motivated by the succcess of density-functional theory in materials description, we first introduce the single particle view on quantum transport that is most easily coupled to density-functional theory. Next, in Sec. 3, we present the educational model of tunneling through a rectangular barrier and introduce the concept of resonant tunneling. Sec. 4 follows with the Landauer approach to linear-response transport theory, where the transmission probability is directly related to the conductance. Then, in Sec 5 we discuss Bardeen's model of tunneling. In the same section, the very important concept of the complex band structure is introduced, providing a way to evaluate the decay parameter in insulators. Finally, Sec. 7 is devoted to applications with technological relevance.

2 Single Particle View on Quantum Transport

Even when restricting the theoretical description to the electron system only, the transport process is actually a complicated many electron problem of a system in non-equilibrium. On the macroscopic scale one can already define some of the different quantities describing the system like current and charge density or the applied electric field. However, it is very difficult to track these quantities down to the microscopic scale due to the complicated thermodynamically averaging taking place. In the following we will only deal with the very restricted subset of phenomena that are due to quantum mechanical nature of the electrons and we will therefore consider systems which can be described by pure wavefunctions without any statistical averag-

[2] With the exception of STM/STS, where the barrier thickness is controlled during experiment.

ing. Still, in this picture one would have to describe the electron transport by the time-dependent many-body wavefunction $\Psi(t)$. For example one could consider the probability $P_{i,f}$ of the system changing its state from some initial multi-electron state Ψ_i into a final multi-electron state Ψ_f where the two states differ with respect to their charge distribution. Thus, this approach makes it necessary to calculate the many particle time-dependent wavefunctions of the entire system. This is a very difficult task which cannot be solved in general.

To overcome this fundamental obstacle we will switch to the single-electron picture of electron transport. Similar to the replacement of the many particle problem of determining the ground state of a many-body quantum mechanical system in density functional theory by a single particle Kohn-Sham formalism we will treat the electronic transport as due to the transport of many independent single-electrons. Furthermore, we will assume that the single particle states in the Kohn-Sham formulation actually describe these independent single-electrons. Of course, there can be no hope that this very simplistic model actually is able to catch all the essential physics of the transport process. However, in analogy to standard band-structure calculations of solids in equilibrium in which in many cases the Kohn-Sham single particle eigenvalues and eigenstates can be successfully interpreted as the elementary excitations of the systems we will apply the same procedure to the electronic transport and assume that the effects of the atomic arrangement, of the electronic (self-consistent) charge density and single particle potential on the current can actually be modeled by this approach.

Many effects restrict the validity of the single-electron approach. Most obvious might be electron-electron scattering effects of different conducting electrons, but also interactions with the lattice beyond the static approximation, i.e. electron-phonon scattering, screening and charging effects or many particle interactions in magnetic systems might limit the validity of the single electron picture. Only if these processes are sufficiently weak, one can hope that the single particle approximations in terms of the Kohn-Sham states will provide reasonable results. This corresponds to the limit dicussed in the introduction in which the mean free path is much larger than the system size. In the single particle picture the description in terms of the Kohn-Sham wavefunction will hold only on length scales shorter than this length scale, since these processes not included in the model will lead to the scattering destroying the phase coherence between the single-electron states involved.

One should note, that in many cases this description of the electron current in terms of single particle physics is a completely inappropriate point of view. As soon as quantum many-body effects come into play, qualitatively different phenomena can be observed. Examples of such effects are the Kondo-effect in which a two level quantum scatterer embedded in a metallic environment leads to conductance abnormalities at low temperatures or correlation effects like the Coulomb blockade which are not reproduced in standard density functional theory treatments.

Obviously our description does not include any processes by which the single particle energy of the states carrying the current is changed. Hence only elastic transport can be described. This kind of quantum transport with only elastic scattering included is frequently called ballistic transport in mesoscopic physics as in many aspects the electrons behave like classical particles moving in a “billiard” like fashion. However, one should be aware of underlying quantum mechanical picture of single electron states by which the electrons are described. These single particle states describe the movement of the electrons between scattering event which scatter them from one single particle state into another.

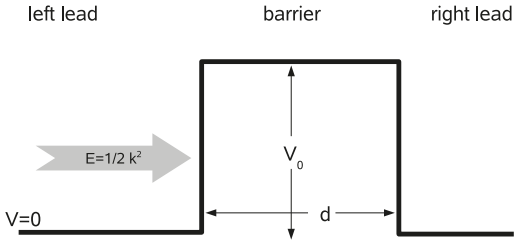


Fig. 1: *Simple one-dimensional model for quantum mechanical tunneling. A electron of energy* $E = \frac{\hbar^2}{2m}k^2$ *is incident from the left to a rectangular barrier potential.*

3 Tunneling through a rectangular Barrier

Of course the quantum mechanical tunneling effect is a very basic phenomenon discussed in every introductory course of quantum mechanics. In brief it describes the fact that in contrast to the classical mechanics which prohibits a particle to enter any area in which the potential level surpasses the particles energy, quantum mechanics assigns a finite non-zero probability find the particle in such areas. Mathematically, this is reflected by the fact that the wavefunction is non-zero in such areas of a repulsive potential. As a direct consequence a particle – in our case an electron – can overcome potential barriers and "tunnel" through regions of space which are classically forbidden to access.

3.1 Analytical solution for the rectangular barrier model

To elucidate this effect a little more, let us consider the probably simplest model for electronic tunneling.
Fig. 1 shows the setup chosen for this simple model, a rectangular barrier of height V_0 and width d between leads or electrodes in which the electrons are described by free electron wavefunctions. In this system it is an trivial problem to construct the wavefunction as

$$\psi(x) = \begin{cases} \exp(ikx) + r\,\exp(-ikx) & x < 0 \text{ in left region} \\ a\,\exp(-\kappa x) + b\,\exp(\kappa x) & 0 < x < d \text{ in the barrier} \\ t\,\exp(ikx) & x > d \text{ in right region.} \end{cases} \tag{1}$$

The wavenumber is given by $k(E) = \sqrt{\frac{2mE}{\hbar^2}E}$, the decay constant by $\kappa(E) = \sqrt{\frac{2m}{\hbar^2}(V_0 - E)}$, the coefficients a, b and r, t can be determined by wavefunction matching, i.e. by the requirement that the wavefunction and its derivative are continuous at $x = 0$ and $x = d$. Simple algebra reveals the well known formula

$$t = \frac{4i\kappa k\, e^{-ikd}}{(ik+\kappa)^2 e^{-\kappa d} + (k+i\kappa)^2 e^{\kappa d}}. \tag{2}$$

For the case of a sufficiently thick and/or high barrier, i.e. large d and/or large κ this expression for t can be simplified by neglecting higher order terms in $e^{-\kappa d}$ to

$$t \sim \frac{4i\kappa k\, e^{-ikd}}{(k+i\kappa)^2} e^{-\kappa d}.$$

The wavefunction now actually leads to an electric current flowing across the barrier. Applying the quantum mechanical current operator one obtains for the current density (which can be most simple evaluated in the right electrode region but is of course conserved in all space)

$$j(x) \propto \frac{1}{2i}\left(\psi^*(x)\partial_x\psi(x) - \partial_x\psi^*(x)\psi(x)\right) \propto |t|^2. \tag{3}$$

Hence we find that there is a finite electric current flowing through the barrier which is proportional to the square of the so called transmission amplitude t, a quantity that can be interpreted as a transmission probability and takes the following elegant analytical form:

$$|t(E)|^2 = \left[1 + \left(\frac{k}{\kappa} + \frac{\kappa}{k}\right)^2 \frac{\sinh^2(\kappa d)}{4}\right]^{-1} \tag{4}$$

for $E < V_0$ and its analytical continuation by defining $\kappa(E) = iq(E) = i\sqrt{\frac{\hbar^2}{2m}(E - V_0)}$ for $E > V_0$:

$$|t(E)|^2 = \left[1 + \left(\frac{k}{q} - \frac{q}{k}\right)^2 \frac{\sin^2(qd)}{4}\right]^{-1}. \tag{5}$$

3.2 Resonant tunneling through virtual bound states

In many cases, the barrier may include defects that produce bound states with energy E_b. These states may interact weakly with the lead wave functions, as both penetrate in the barrier region with exponentially decaying amplitude. The interaction gives to the bound states a small energy broadening and a finite lifetime, where an electron may hop out of the bound state to the leads. One then speaks of a *virtual bound state* or a *resonant state*. Under this condition, the transmission probability between the leads can reach unity at an energy $E_r \approx E_b$, while it remains very small at different energies.

The situation can be modelled by a rectangular barrier with an attractive δ-potential in the middle that represents the defect. The potential reads

$$V(x) = \begin{cases} V_0 - \lambda\delta(x), & -d/2 < x < d/2 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

and is schematically shown in Fig. 2 (top). For the wavefunction we make the *Ansatz*

$$\psi(x) = \begin{cases} \exp(ikx) + r\exp(-ikx), & x < -d/2 \\ a\exp(-\kappa x) + b\exp(\kappa x), & -d/2 \le x < 0 \\ c\exp(-\kappa x) + d\exp(\kappa x), & 0 \le x < d/2 \\ t\exp(ikx), & d/2 \le x \end{cases} \tag{7}$$

in analogy to Eq. (1). Again, the wavenumbers k and $\kappa = \sqrt{\frac{2m}{\hbar^2}(V_0 - E)}$ are energy-dependent quantities and both are real-valued for the tunneling case ($E < V_0$). At $E > V_0$ we may exchange κ with $q = \sqrt{\frac{2m}{\hbar^2}(E - V_0)}$, or simply allow κ to take imaginary values. The six coefficients (r, a, b, c, d, t) are found by matching the boundary conditions at the points of discontinuity of the potential. However, while at $x = \pm d/2$ the boundary conditions correspond to

continuity of both the wavefunction and its derivative, at $x = 0$ only the wavefunction is continuous, while its derivative ψ' shows a finite-size step due to the δ-function potential. One can recognise this by the following standard manipulation given in quantum mechanics textbooks. Integrating the Schrödinger equation, $-\frac{\hbar^2}{2m}\psi''(x) - \lambda\delta(x)\psi(x) = E\psi(x)$, in a small interval $[-\delta, \delta]$ around $x = 0$, and letting $\delta \to 0$, the right-hand-side $E\int_{-\delta}^{\delta}\psi(x)dx$ vanishes because ψ is continuous, while the left-hand side gives $-\frac{\hbar^2}{2m}[\psi'(0^+) - \psi'(0^-)] - \lambda\psi(0) = 0$. This is the finite-step condition for ψ'.

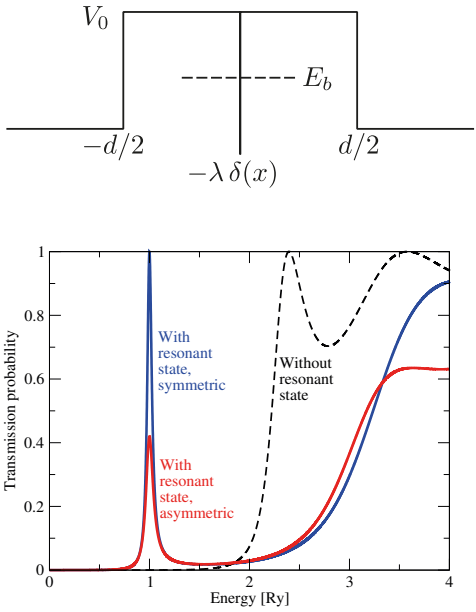


Fig. 2: *Top: A rectangular barrier of height V_0 and width d with an attractive δ-potential of the form $-\lambda\,\delta(x)$ in the middle. The strength λ is chosen such that produces a virtual bound state at $E_b > 0$. Bottom: Blue curve: Transmission probability for such a barrier as a function of energy. The parameters used are $d = 5a_B$, $V_0 = 2$Ry, and $\lambda = 2$Rya_B, leading to $E_b = 1$Ry. The tunneling probability (at $E < V_0$) peaks to full (unitary) transmission very close to E_b. Red curve: The same but with an asymmetric position of the δ-potential, $-\lambda\delta(x - x_0)$ with $x_0 = 0.5a_B$. The transmission peak does not reach unity. Black, dashed curve: transmission probability of the same barrier but without the attractive δ-potential.*

It is straightforward to solve the resulting 6×6 system of equations for (r, a, b, c, d, t) analytically, because it has a band-diagonal form boiling down to a recursive solution of 2×2 systems. However, the resulting expression is rather complicated. Here, we present in Fig. 2 only the graph of the end-solution for a special case of $d = 5a_B$ and $V_0 = 2$Ry.[3] Three cases for the transmission probability $|t(E)|^2$ are shown.

[3]One Bohr radius is $a_B = 0.529177$Å $= 0.529177 \times 10^{-10}$m. One Rydberg is 1Ry $= 13.6058$eV and corresponds to the excitation energy of the Hydrogen atom.

First, the black, dashed curve shows the case without an attractive δ-potential ($\lambda = 0$), i.e., the result of Eqs. (4,5). For $E < V_0$ we have $|t(E)|^2 < 1$, exponentially decaying at lower energies, as Eq. (4) demands. For $E > V_0$ (i.e., outside the tunneling condition) the transmission probability shows an oscillatory behavior with resonances, even reaching unity; at $E \gg V_0$ we have $q/k \approx 1$ and Eq. (5) thus gives $|t(E)|^2 \to 1$.

Second, the blue continuous curve shows $|t(E)|^2$ in the presence of an attractive δ-potential [Eq. (6)] with $\lambda = 2\text{Ry}a_B$. The curve looks overall different than the case $\lambda = 0$, but the most striking result is the sharp transmission peak at $E = E_r \approx 1\text{Ry}$, with $|t(E_r)|^2 = 1$. The full transmission here is caused by the virtual bound state (discussed in beginning of the present section) at $E_r \approx E_b$, where $E_b = 1\text{Ry}$ is the level of the bound state that would exist if the barrier limits would extend to $\pm\infty$.[4] The reason that $E_r \approx E_b$ and not $E_r = E_b$ is that the hybridization of the bound state with the continuum is asymmetric in energy, shifting E_r slightly lower than E_b; but this small effect is smaller than the line thickness of the present plot.

Third, the red continuous curve demonstrates the case that the potential is *not symmetric*; in particular the δ-potential was shifted to the positive x-axis, reading $-\lambda\delta(x - x_0)$ with $x_0 = 0.5a_B$. Here, the transmission peak at E_r is also present but does not reach unity any more. Thus we see that the presence and energy of a transmission resonance follows from the virtual bound state, but the transmission peak value strongly depends on the spatial position of the virtual bound state; the highest value is reached in the fully symmetric case.

3.3 Resonant tunneling through surface states

A more complicated form of resonant tunneling was discovered by numerical, density-functional calculations and explained by an analytical model by Wunnicke and co-workers [4]. Here we summarise their findings. When investigating the transmission in insulating tunnel barriers separating metallic surfaces, the authors found that there exist *tunneling hot spots* in the (two-dimensional) surface Brillouin zone, i.e., small regions where the transmission peaks and reaches unity, while it is very small in the overwhelmingly larger part of the surface Brilloun zone. An analysis of the energy-resolved density of states at the interface showed that the hot-spot positions coincide with positions of resonant states at the metal/insulator interface. For the occurrence of the effect it is important that the setup is symmetric, i.e., that there are two interface states contributing to the transmission peak, one at each interface, at the same position in the surface Brillouin zone and at the same energy. The two interact with each other and produce bonding and antibonding hybrids with a splitting of ΔE, while each of them also interacts with the continuum of metal states and obtains an energy spread Γ. As long as $\Delta E > \Gamma$, unitary transmission is maintained; but as the barrier thickness grows, the split ΔE decreases, and eventually $\Delta E < \Gamma$, after which point the transmission drops exponentially with thickness.

The effect reminds of the mechanism described in the previous subsection, with the difference that the role of defects is played by the interface resonant states, and that there are two such states (one from each interface) contributing to the transmission peak.

[4] The attractive potential $V(x) = -\lambda\delta(x)$, $\lambda > 0$, produces a bound state at $E_b = -m\lambda^2/(2\hbar^2)$. Here, the virtual bound state position in the barrier is found by shifting this value by V_0.

4 Landauer formula

Landauer [5] proposed a theory of the transport process which is well adapted to describe the tunneling transport. A very intuitive and simple derivation will be presented here. The Landauer equation can also be derived more rigorously starting from linear response theory. In the Landauer approach to transport one considers the region Ω – in which the electrons travel ballistically – to be attached to two reservoirs L and R.
The conductance Γ of the region Ω is defined by the current I_{LR} divided by the potential difference between the two reservoirs. The current I_{LR} on the other hand is given by the current due to all electrons traveling from L to R minus the current due to the electrons traveling vice versa

$$I_{LR} = I_{L\to R} - I_{R\to L}. \tag{8}$$

To arrive at an equation for these currents, one can start with a simple one-dimensional model. The current from the left to the right is determined by all electrons leaving the left reservoir, entering the scattering region Ω, and leaving this scattering region by passing into the right reservoir. If one now assumes a very simple picture of the region Ω in which its electronic structure is described by single band in which states with $k > 0$ propagate from the left to the right the current is given by an integral over all states with $k > 0$ up to the Fermi wave-vector k_F

$$I_{L\to R} = \int_0^{k_F} ev(k)dk, \tag{9}$$

where v denotes the group velocity of the state. Since

$$v = \frac{1}{\hbar}\frac{\partial E}{\partial k} \tag{10}$$

and converting the integral over k into an energy integration using the density of states $n(E)$,

$$\begin{aligned} I_{L\to R} &= \int_0^{\mu_L} \frac{e}{\hbar}\frac{\partial E}{\partial k} n(E)\, dE \\ &= \int_0^{\mu_L} \frac{e}{\hbar}\frac{\partial E}{\partial k}\,\frac{1}{\partial E/\partial k}\,\frac{1}{2\pi}\, dE \\ &= \int_0^{\mu_L} \frac{e}{h}\, dE = \frac{e}{h}\,\mu_L\,, \end{aligned} \tag{11}$$

where the energy integration has to be performed over all energies up to the Fermi energy (the chemical potential) of the left reservoir. This can be understood from the requirement that the electrons were assumed to be incoming from the left and therefore must be occupied in the reservoir.
Using the same derivation for the states incoming from the right reservoir one obtains

$$I_{LR} = \frac{e}{h}(\mu_L - \mu_R). \tag{12}$$

Identifying the difference in the chemical potentials μ_L and μ_R with the applied voltage $eV = (\mu_L - \mu_R)$ one obtains the following interesting equation for the conductance

$$\Gamma = \frac{I_{LR}}{V} = \frac{e^2}{h}. \tag{13}$$

This equation is truly remarkable since it states that each conducting band contributes the same to the conductance. Irrespectively of the density of states or the group velocity of the conducting states the conductance is always given by the fundamental quantum of conductance $\frac{e^2}{h}$. Indeed, as Eq. (10) shows, states with a low velocity and therefore a low current $j = ev$ are compensated by their higher density of states such that the conductance remains constant.
In the case of multiple bands, the derivation has to be modified by the inclusion of an extra sum over the different bands. Therefore in the general case of N conducting bands one obtains

$$\Gamma = \frac{e^2}{h} N. \tag{14}$$

The different „bands“ in this context are usually called „channels“. The argumentation presented so far did not care about the proper definition of these channels. These were simple assumed to form some kind of „band“ within Ω described by the usual formalism of a wave-vector k and a dispersion relation $E(k)$. Strictly speaking, since the system is not periodic, one cannot speak of Bloch states with some wave-vector having a component k in the direction of the current.
Since a key point in the discussion was the preparation of a state traveling from within the reservoirs through the region Ω one should clarify this idea. For such a state traveling to the right, one might assume the typical scattering problem. Within the left reservoir one considers a wavefunction being a Bloch state propagating towards the region Ω. „Propagate towards“ in this context should be understood as a state having a current flowing towards Ω. Within the reservoirs the resulting scattering state can be written in terms of reflected ψ_r and transmitted ψ_t states which are all solutions of the bulk Schr$\acute{}$odinger equation in the reservoirs with the same energy as the incoming state ψ_{in} . The k values of these transmitted and reflected states have to be chosen such that the states „propagate away“ from Ω.

$$\psi(\vec{r}) = \begin{cases} \psi_{\text{in}}(\vec{r}) + \sum_n r_{\text{in},n}\, \psi_r^n(\vec{r}) & \vec{r}\,\text{in left reservoir} \\ \sum_{n'} t_{\text{in},n'}\, \psi_t^{n'}(\vec{r}) & \vec{r}\,\text{in right reservoir} \end{cases} . \tag{15}$$

Here, the summations can be considered to be performed over all reflected Bloch states or all transmitted Bloch states. In principle, also states decaying away from the interfaces into the reservoirs must be included in this expansion. However, since these do not carry any current and by shifting the interface far enough into the reservoirs one can eliminate these decaying states.
Looking back to the derivation of the Landauer formula, an important change has to be made. While in Eq. (9) and Eq. (10) the summation over the incoming states and the evaluation of the current from their group velocities were all performed within the same single band picture, now one has to distinguish more carefully. The k integration in Eq. (9) has to be performed over the „in“ label of the expansion in Eq. (15). The sum over the velocities on the other hand is best performed in the right electrode. This is possible since current is conserved and can be very easily be done if all transmitted states and the incoming state are normalized to carry unit current. Using the orthogonality of the Bloch states one can perform the same steps as in Eq. (8) to (12) again to derive the more general Landauer equation for ballistic transport in the presence of some scattering of the incoming electrons,

$$\Gamma = \frac{e^2}{h} \sum |t_{ij}|^2, \tag{16}$$

where i, j label the Bloch states in the reservoirs traveling from the left to the right.

Eq. (16) allows a simple interpretation of the transport in terms of the underlying quantum mechanical property of the transmission probability $P_{ij} = |t_{ij}|^2$ of an electron from the incoming Bloch state i into the transmitted Bloch state j. This interpretation makes the requirement of normalizing the incoming and transmitted Bloch states to unit current very clear, since in this normalization the direct interpretation of this probability is reasonably well defined and Eq. 16 can be seen as a simple generalization of Eq. 14.

4.1 Interpretation of the Landauer formula

The Landauer formula Eq. (16) was the source of some confusion for quite some time after its first formulation [6, 7]. The most striking feature of the equation might be its limit for a perfectly transmitting region, i.e. for a region with $P_{ij} = |t_{ij}|^2 = 1$ for some set of i, j. For example if one would consider a perfect bulk crystal sandwiched between reservoirs of the same bulk material the expansion of Eq. (15) would collapse to

$$\psi(\vec{r}) = \begin{cases} \psi_{\text{in}}(\vec{r}) + \sum 0\psi_r & \vec{r} \text{ in left reservoir} \\ 1\,\psi_t = \psi_{\text{in}}(\vec{r}) & \vec{r} \text{ in right reservoir} \end{cases}, \tag{17}$$

and one would rediscover Eq. (14) with N denoting the number of incoming Bloch states. At first glance, this means that the Landauer equation predicts a limited conductance of a system without any de-coherent scattering, i.e. of a perfect bulk crystal. In the same way the Landauer equation would also give a finite conductivity of a free electron gas. In this case a question which can always be asked only becomes more obvious to ask: How can a region with ballistic transport, i.e. without any dissipative processes, have a finite conductance? Since there is a voltage drop over the region and a current is flowing, some energy must dissipate. The key to the answer to this question lies in the definition of the reservoirs which were assumed to be in thermal equilibrium with some chemical potential μ attached to them. This is only possible, if there are actually dissipative processes in the reservoirs leading to the „thermalization“ of the „hot“ electrons being transfered across the region of ballistic transport.
The surprising result of a finite conductance in the case of a perfect crystal can now been interpreted in different ways. Either the setup described was not correct, since the reservoirs could not remain in thermal equilibrium and being perfect crystals like the region of ballistic transport at the same time or, which is actually very much the same, no finite voltage can be applied across such a system. The finite conductance of such a system with perfect ballistic transmission can now be interpreted as due the finite resistance at the interface between the reservoir and the ballistic region. This is also called the Sharvin-resistance of the system.
Another point to mention in the discussion of the physical significance of the Landauer equation is its formulation in terms of a two-terminal device. Both the current and the voltage drop are defined between the same two reservoirs. In many experiments a four point measurement is performed in which the current is driven between electrodes different than those between the voltage drop is measured. B$\acute{}$uttiker [6] presented a generalization of the Landauer equation to these multi-terminal case. While this approach is very appropriate for mesoscopic physics, on the atomic scale multi-terminal arrangements are not the typical experimental arrangement and thus Eq. (16) will be sufficient. Additional resistances present in the current circuit are frequently eliminated in a four-point measurement, in which two additional potential probes are attached close to the scattering volume. However, for scattering volumes on the atomic scale, these geometries are not appropriate and thus we will restrict ourself to simple two point geometries.

While the Landauer equation is valid in many cases reaching from systems with high conductivity to systems in the tunneling regime, one has to be careful in its application in some cases. Only states in which the incoming and transmitted waves can be described by Bloch states contribute to the tunneling current. This excludes any state which is localized within the region of ballistic transport to contribute. This corresponds to the fact that these states do not carry any current within the simple one electron picture of transport chosen. In reality, there exist processes beyond this picture which lead to some coupling of these localized states to the otherwise orthogonal Bloch states in the reservoirs. For example the many-body electron-electron interaction, electron-phonon scattering, or structural defects not included in the description can provide such a coupling. Thus, while the Landauer approach will be correct for cases of high transmission through Bloch states, one could imagine that in the limit of a very low transmission probability another processes of transport across the ballistic region becomes important. In the one electron picture these processes could be thought of as the transition of an electron from the reservoirs into some localized state of the reservoir, the transition of the electron from one side of the reservoir to the other and than the transition of the electron into a state of the other reservoir. The validity of the Landauer model is now limited by the transmission probability between the reservoir states and the localized state. If this probability becomes comparable to the probabilities $P_{ij} = |t_{ij}|^2$ the Landauer equation breaks down.
On the other hand, one can of course treat the other limit in which the transition probability between the two sides of the reservoir becomes very small and the details of the scattering processes needed to couple the states can be neglected. This limit can be successfully described by theories for the quantum mechanical tunneling process.

5 The Bardeen Approach to Tunneling

The following description of the tunneling process is based on Bardeen's approach to tunneling which essentially applies time dependent perturbation theory to the problem. Fig. 3 shows the tunneling setup used in this approach. Two semi-infinite crystals are separated by a barrier region, which will be assumed to be a vacuum barrier for simplicity. If this vacuum barrier is sufficiently high and wide one can think the total setup to consist of of two independent systems: one at the left (L) and one at the right(R) side.

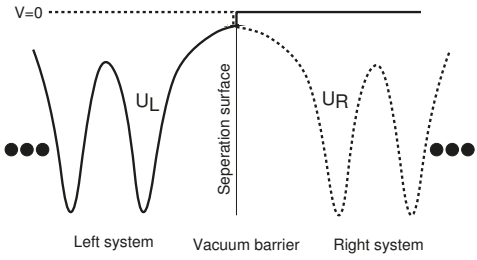


Fig. 3: *Tunneling setup used in Bardeen's approach to transport. The two semi-infinite crystals at the left and the right are separated by a vacuum barrier.*

This total separation of the systems leads to two independent Schrödinger equations for the two

sides

$$(T + U_L)\psi_L = \epsilon_L \psi_L$$

$$(T + U_R)\psi_R = \epsilon_R \psi_R \tag{18}$$

where T denotes the operator of the kinetic energy of a single electron and U_L and U_R are the potentials of the left and right system respectively. The single particle wavefunction $\psi(t)$ of the entire setup is determined by the total Hamiltonian $H = T + U_L + U_R$.

Now one can apply time dependent perturbation theory to describe the tunneling of an electron across the vacuum barrier. Tunneling from the left to the right is assumed, the case of an electron tunneling vice versa may be treated completely analogously. The initial state of the tunneling process is localized in the left system. Therefore, there exists an eigenstate ψ_L^μ with $|\psi(t \to -\infty)\rangle = |\psi_L^\mu\rangle$. The time dependence of the state $|\Psi(t)\rangle$ is governed by the Hamiltonian of the whole system.

$$i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = H |\psi(t)\rangle . \tag{19}$$

The tunneling probability is given by the overlap of this time-dependent wavefunction with a wavefunction $|\psi_R^\nu >$ of the right system. Multiplying Eq. (19) from the left with $\langle\psi_R^\nu|$leads to

$$\langle\psi_R^\nu| \left(i\hbar \frac{\partial}{\partial t} \right) |\psi(t)\rangle = \langle\psi_R^\nu| H |\psi(t)\rangle . \tag{20}$$

Using the Schrödinger equation for the left state one obtains

$$\begin{aligned} i\hbar \tfrac{\partial}{\partial t} \langle\psi_R^\nu|\, \psi\rangle &= \langle\psi_R^\nu| H |\psi\rangle - \langle\psi_R^\nu| H_R |\psi\rangle \\ &= \langle\psi_R^\nu| U_L |\psi\rangle \end{aligned} \tag{21}$$

Substituting $|\psi(t \to -\infty)\rangle = |\psi_L^\mu\rangle$ for $|\psi\rangle$ at the right hand side of Eq. (21) leads to first order perturbation theory

$$i\hbar \frac{\partial}{\partial t} \langle\psi_R^\nu|\, \psi\rangle = \langle\psi_R^\nu| U_L |\psi_L^\mu\rangle . \tag{22}$$

Even though this equation looks familiar one has to emphasize that this is not a result obtained by standard time-dependent perturbation theory. The states $|\psi_L\rangle$ and $|\psi_R\rangle$ are eigenstates of the Hamiltonians H_L and H_R respectively. Therefore, they do not form a complete orthogonal basis of the eigenspace of the total Hamiltonian $H = T + U_L + U_R$ and the matrix elements at the left side of Eq. (22) are not sufficient to determine the total time dependence of $|\psi >$. This is a basic weakness of Bardeen's approach. However many applications [8, 9, 10, 11] of this formalism have shown that Bardeen's approximation produces reliable results for systems which are well separated, i.e. systems where the overlap of the two wavefunctions ψ_R and ψ_L is small.

Since the potential U_L is not small in the left region, the question arises whether one is allowed to use perturbation theory at all. However, it can be seen from Eq. (22) that the quantity which in fact determines the strength of the perturbation of the initial state is $\langle\psi_R^\nu| U_L |\psi_L^\mu\rangle$. Since the final wavefunction $|\psi_R\rangle$ is localized in the right region in which the left potential U_L is very weak this perturbation might still be regarded as a small perturbation and thus time depended perturbation will lead to reasonable results.

By separating the time-dependence of the states $|\psi_L^\mu\rangle = e^{i\epsilon_\mu t}\,|\Psi_L^\mu\rangle$ and $|\psi_R^\nu\rangle = e^{i\epsilon_\nu t}\,|\Psi_R^\nu\rangle$, integrating Eq. (22) and performing the limit $t \to \infty$, one obtains an expression for the tunneling-probability per time interval

$$P_{\mu\nu}^{LR} = \lim_{t\to\infty} \frac{1}{t}\frac{1}{\hbar^2}\int_0^t |\langle\psi_R^\nu|\,U_L\,|\psi_L^\mu\rangle|^2 \tag{23}$$

$$= \lim_{t\to\infty} \frac{4}{\hbar}\frac{\sin^2\left(\frac{\epsilon_\nu-\epsilon_\mu}{2\hbar}t\right)}{(\epsilon_\nu-\epsilon_\mu)^2 t}\left|M_{\mu\nu}^{RL}\right|^2, \tag{24}$$

where the matrix element $M_{\mu\nu}^{LR}$ is given by the stationary-state matrix element of the potential

$$M_{\mu\nu}^{LR} = \langle\psi_R^\nu|\,U_L\,|\psi_L^\mu\rangle\,. \tag{25}$$

Assuming a continuous range of energy levels ϵ_μ (or ϵ_ν) the limit of Eq. (24) can be evaluated directly [12]. One obtains

$$P_{\mu\nu}^{LR} = \frac{2\pi}{\hbar}\delta(\epsilon_\nu - \epsilon_\mu)|M_{\mu\nu}^{LR}|^2. \tag{26}$$

This result is similar to the well known 'Golden Rule' Fermi obtained for standard time-dependent perturbation theory. It describes elastic tunneling with energy $\epsilon_\nu = \epsilon_\mu$ only. Formally this condition is taken care of by the δ-function in Eq. (26).
To evaluate this matrix element one can introduce an additional approximation. He assumed the potential U_L to be zero in the right region of space. Similar the right potential should be zero in the left region. More formal one assumes a separation surface S which separates the regions in which the two potentials differ from zero. This can be written down by the condition $U_L U_R = 0$ for any point in space. Figure 4 shows the setup as used in this additional approximation.

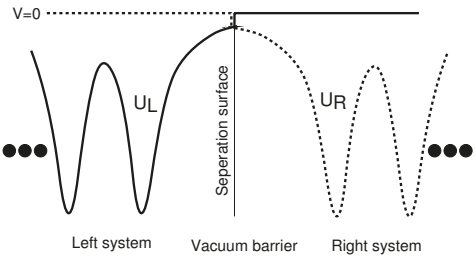


Fig. 4: *Potential used in the Bardeen approach to tunneling. The left (right) potential U_L (U_R) is then assumed to be zero in the right (left) region.*

Of course, this approximation will become better if the potentials U_L and U_R are reasonably small at and beyond the separation surface. This will be the case if the separation surface is located far out in the vacuum.
Using the Schrödinger equation for the left wavefunction and having in mind that the potential U_L is zero in the right space one can now rewrite the matrix element as an integral over the left region only

$$M_{\mu\nu}^{LR} = \int_L \Psi_R^\nu(\vec{r})^*(\epsilon_\mu + \frac{\hbar^2}{2m}\vec{\nabla}^2)\Psi_L^\mu(\vec{r})dV \tag{27}$$

which can be written in a more symmetric form

$$\begin{aligned} M_{\mu\nu}^{RL} &= \int_L \left\{ \Psi_R^\nu(\vec{r})^* \epsilon_\nu \Psi_L^\mu(\vec{r}) + \Psi_R^\nu(\vec{r})^* \tfrac{\hbar^2}{2m} \vec{\nabla}^2 \Psi_L^\mu(\vec{r}) \right\} dV \\ &= \int_L \left\{ \Psi_R^\nu(\vec{r})^* \overleftarrow{(T+U_R)} \Psi_L^\mu(\vec{r}) + \Psi_R^\nu(\vec{r})^* \tfrac{\hbar^2}{2m} \vec{\nabla}^2 \Psi_L^\mu(\vec{r}) \right\} dV \\ &= -\tfrac{\hbar^2}{2m} \int_L \left\{ \Psi_L^\mu(\vec{r}) \vec{\nabla}^2 \Psi_R^\nu(\vec{r})^* - \Psi_R^\nu(\vec{r})^* \vec{\nabla}^2 \Psi_L^\mu(\vec{r}) \right\} dV \end{aligned} \tag{28}$$

In these transformations in the first step the eigenvalue ϵ_μ was substituted by ϵ_ν because energy conservation requires the calculation of matrix elements with $\epsilon_\mu = \epsilon_\nu$ only. In the second step the Schrödinger equation for the right state was used (the arrow indicates the wavefunction the operators acts on). The integration area is the left region. Since the potential U_R is assumed to be zero in this region, it was dropped in the last step. Using Greens theorem and the boundary condition that the right wavefunction is zero at infinite distance from the separation surface this integral can be transformed into an integral over the separation surface

$$M_{\mu\nu}^{LR} = -\frac{\hbar^2}{2m} \int_S \left(\Psi_L^\mu(\vec{r}) \vec{\nabla} \Psi_R^\nu(\vec{r})^* - \Psi_R^\nu(\vec{r})^* \vec{\nabla} \Psi_L^\mu(\vec{r}) \right) dS. \tag{29}$$

So far only an expression for the probability of the transition of an electron from a left state into a right state was obtained.
Slightly modifying Eq. (26) this probability can be written as

$$P_{\mu\nu}^{LR} = \frac{2\pi}{\hbar} \delta(\epsilon_\mu^L - \epsilon_\nu^R - eV) |M_{\mu\nu}^{LR}|^2, \tag{30}$$

where the additional term eV is introduced to account for the bias voltage V applied between the two sides. To calculate the tunneling current one has to sum over all different possible left and right states and one has to keep in mind that the electrons might tunnel from the left to the right as well as vice versa. The total current therefore is given by

$$\begin{aligned} I &= I^{L\to R} - I^{R\to L} \\ &= e \sum_{\mu\nu} f(\epsilon_\mu)(1 - f(\epsilon_\nu + eV)) P_{\mu\nu}^{LR} - e \sum_{\mu\nu} (1 - f(\epsilon_\mu)) f(\epsilon_\nu + eV) P_{\nu\mu}^{RL} \\ &= e \sum_{\mu\nu} \left(f(\epsilon_\mu) - f(\epsilon_\nu + eV) \right) P_{\mu\nu}^{LR} \end{aligned} \tag{31}$$

where $f(\epsilon)$ denotes the Fermi-distribution function which is introduced to ensure that only tunneling from occupied to unoccupied states can occur. In Eq. (31) the symmetry of the tunneling probability $P_{\mu\nu}^{LR} = P_{\nu\mu}^{RL}$ which can easily be deduced from Eq. (29) was used. The sum in Eq. (31) has to be performed over all right states labeled by ν and all left states labeled by μ. No further assumption is made on the nature of these left and right states, i.e. both Bloch states and surface states decaying into the bulk contribute to the current and therefor this formula differs significantly from the Landauer formula.

5.1 Landauer conductance versus Bardeen's tunneling

One might wonder which the difference will be between the results of the Bardeen formula of tunneling and of the Landauer approach. Of course, as stressed in Sec. 5 the results will differ

significantly as soon as localized states are present in the vicinity of the barrier region. However, it remains to be clarified what should be expected in the absence of these states. Before the difference between these two transport formulas will be investigated for more realistic systems, it is instructive to look back at the very simple model of a rectangular barrier we discussed in the beginning.

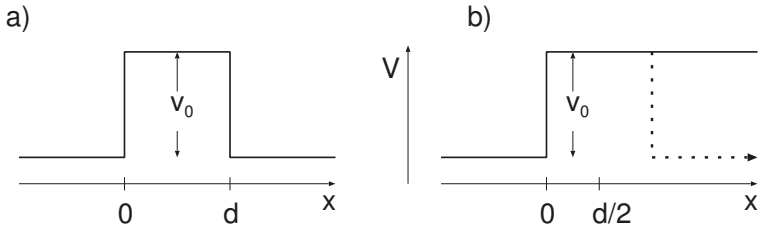


Fig. 5: *a) Setup of a one-dimensional rectangular barrier as discussed before. b) Corresponding system for the Bardeen formula. The two sides are separated by extending the Barrier to infinity. Only the left system of Bardeen's setup is shown, the right is constructed mirrored.*

Inserting the expression for the transmission into the Landauer equation (Eq. (16)) one obtains the tunneling conductance of this simple barrier in the Landauer approach

$$\begin{aligned} \Gamma_{\mathrm{L}} &= \frac{e^2}{h}\frac{(4\kappa k)^2}{(k^2+\kappa^2)^2(1-e^{-2\kappa d})^2+(4\kappa k)^2 e^{-2\kappa d}} e^{-2\kappa d} \\ &= \frac{e^2}{h}\left|\frac{4\kappa k}{k^2+\kappa^2}\right|^2 e^{-2\kappa d} + O\left(e^{-4\kappa d}\right). \end{aligned} \tag{32}$$

As discussed, the second term in this expressions denotes contributions of order $e^{-4\kappa d}$ which can be neglected for any sufficiently thick barrier.
If one wants to treat the same system using Bardeen's formula, one has to separate the two systems by extending the barrier to infinity as indicated in Fig. 5b). The wavefunctions of the two systems are of course equal except for the transformation $x \leftrightarrow d - x$ and can be written, e.g. for the left side, as

$$\psi_L = \begin{cases} \exp(ikx) + \frac{ik+\kappa}{ik-\kappa}\exp(-ikx) & x \text{ in the leads} \\ \frac{2ik}{ik-\kappa}\exp(-\kappa x) & x \text{ in the barrier.} \end{cases} \tag{33}$$

This formula does not depend on the barrier thickness as they describe completely decoupled systems. The conductance in Bardeen's approach is now given by inserting a separation surface at $x = d/2$ and to evaluate the transition probabilities of Eq. (30) at energy ϵ_0 to obtain

$$\begin{aligned} P_{\text{Bardeen}} &= \frac{\pi\hbar^3}{2m^2}\left|2\frac{2ik}{ik-\kappa}e^{-\kappa d/2}\,\kappa\,\frac{2ik}{ik-\kappa}e^{-\kappa d/2}\right|^2 \delta(\epsilon-\epsilon_0) \\ &= \frac{\pi\hbar^3}{2m^2}k^2\left|\frac{4\kappa k}{k^2+\kappa^2}\right|^2 e^{-2\kappa d}\delta(\epsilon-\epsilon_0). \end{aligned} \tag{34}$$

Finally the summation over all states gives according to Eq. (31) gives

$$\Gamma_{\mathrm{B}} = \frac{e^2}{h} \left| \frac{4\kappa k}{k^2 + \kappa^2} \right|^2 e^{-2\kappa d}. \tag{35}$$

Hence, both Eqs. (32) and (35) reveal the same conductance in the limit of large κd, i.e. for a small conductance. Fig. 6 illustrates this equivalence for a set of parameters typical for an STM setup. It is seen that for barriers thicker than $\sim$1 Å Eqs. (35) and (32) give practically identical results.

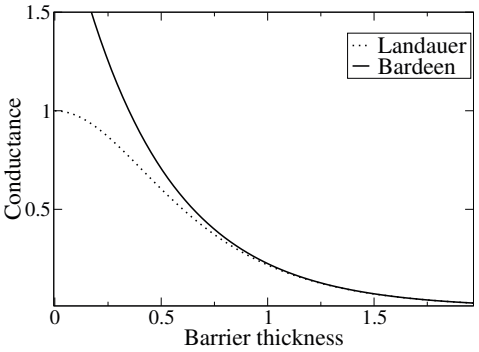


Fig. 6: *Conductance (in units of the conductance quantum e^2/h) through a rectangular barrier of 5eV barrier height between free electron leads ($\epsilon = 1$eV) as a function of the barrier thickness (in Å).*

5.2 Cu-vacuum-Cu tunneling

To demonstrate the two formulas for a more realistic, but still simple case we investigated the electron transmission with normal incidence ($k_\parallel = 0$) through a 8.3 Å vacuum barrier separating two Cu(111) surfaces. The calculation was performed within the DFT using the FLAPW-method, as implemented in the embedding version of the `FLEUR`-code[13, 14, 15].
Figure 7 compares the result of the Landauer formula and of Bardeen's approach. In contrast to the simple analytic model discussed before in this setup both localized surface states and delocalized propagating Bloch states can be found. In the energy range below ≈ -0.9 eV a band of Bloch electrons with $k_\parallel = 0$ can be found which due to their low energies have very low tunneling probabilities. This transmission probability is described by both formulas and similar to the analytic case discussed before both approaches lead to essentially identical results.
At -0.4 eV, however, a surface state on the Cu(111) surface can be found for $k_\parallel = 0$. In the case of the setup required for the Landauer equation these surface states will occur on both sides and hence these states will be split into two levels by hybridization. On the other hand, in Bardeen's approach, because of the complete separation of the two systems only a single surface state peak occurs and no hybridization splitting is seen in Fig. 7. The finite width of this peak is due to the introduction of an imaginary contribution to the energy of $\delta = 0.03$ eV for illustrative purpose only. Such a small broadening of the peak is needed for a numerical integration of the total conductance as well but should not be confused with the modeling of the coupling of the

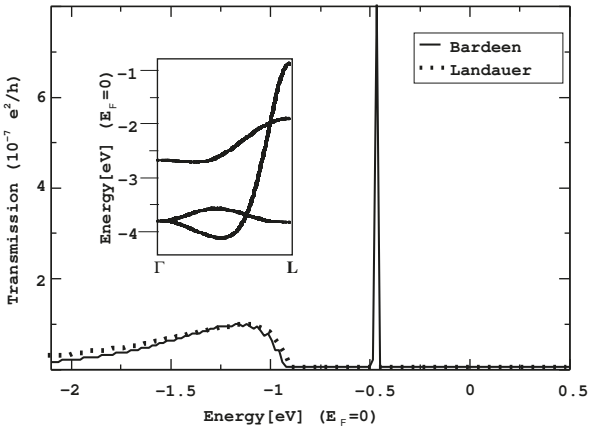


Fig. 7: *Conductance through a vacuum barrier separating two Cu(111) surfaces. Only electrons with normal incidence are considered. The peak width is due to a finite imaginary part in the energy and its height has been scaled down by a factor of five. The inset shows the normal incidence part of the Cu band structure relevant for the energy window in which transmission from Bloch states can be found from both approaches.*

surface state to the leads. For reasonably small imaginary parts the integrated conductance of the peak does not depend on the exact choice of the imaginary energy.

In the case of the setup required for the Landauer equation, one may estimate the electron hopping rate between the two levels from the energy splitting. Nevertheless, these states do not contribute to the conductance in the Landauer equation as no Bloch states are present in the leads at this energy. Bardeen's equation on the other hand contains a strong contribution from the surface state. In this very artificial situation with the same surface state on both sides of the barrier this effect can only be seen at exactly zero bias. However, the arguments presented here are equally valid for situations in which one encounters a localized state at one side of the barrier and a continuum at the other.

In order for electron hopping between two surface states to be measured as the current, they must be coupled to extended states in both leads, for example, via impurity-induced random potentials, electron-phonon interaction,[16] and electron-electron scattering.[17] Applying Bardeen's equation to the surface-state conductance assumes implicitly that the transition between two localized states on both sides is the rate-limiting process.

The present results clearly indicate the limits of validity of the two different approaches. While the Landauer formula will be suitable in cases in which a relatively high conductance is obtained by the coupling of Bloch states, it must be applied with care for tunneling setups in which states localized at interfaces might play a crucial role. On the other hand the Bardeen approach to tunneling is only suitable for exactly these situations and requires a hight barrier leading to a low conductance.

6 Complex band structure

So far we considered simple models only, in which we assumed the electronic wavefunction to either have the form of a plane-wave corresponding to a free particle in a constant potential or to decay exponentionally with a decay constant κ which was directly related to a barrier height V_0as

$$\kappa = \frac{1}{\hbar}\sqrt{2mV_0}. \tag{36}$$

This picture to describe the fundamental properties of the electrons by simple parameters like the barrier height and the effective mass m is of course very intuitive as it can directly be used in an interpretation as presented so far. However, it is known that the underlying assumption of a simple parabolic dispersion of the bands, i.e. the simple quadratic relation between energy and decay constant/momentum, is usually not fulfilled in real materials. While this approximation can be useful as long as one considers processes dominated by electronic states close to band-edges, i.e. at extrema of the bands where a parabolic approximation can be reasonable, tunneling processes often are not in this class of problems. Hence, the description of tunneling in terms of the effective mass of the electrons and a corresponding effective barrier height can be completely misleading.

Instead following the idea of the band structure, in which the relation between crystal momentum k and the energy is specified for the Bloch states, we will discuss the generalization of this concept to the case of complex k-values. The imaginary part of such k values describes the decay of the wavefunctions when translated by a lattice vector $\vec{R}$ according to the Bloch factor

$$e^{i\vec{R}\vec{k}} = e^{i\vec{R}\Re(\vec{k})}e^{-\vec{R}\vec{\kappa}}. \tag{37}$$

In the following we will describe this idea of generalizing the concept of the band structure a bit further.

The usual band structure as described in contribution A2 is of course tightly connected with the Bloch theorem. The Bloch theorem follows from the fact that in an infinite crystal the eigenfunctions of the Hamiltonian can be chosen such that these functions are at the same time eigenfunctions of the translation operator which keeps the lattice invariant. This is the consequence of these translations commuting with the Hamiltonian due to the periodicity of the potential. Hence, we are faced with two eigenvalue problems which have can be solved simultaniously

$$H\psi = \epsilon\psi \tag{38}$$

and

$$T\psi = \lambda\psi \tag{39}$$

in which the eigenvalue of the second equation is interpreted as

$$\lambda = e^{i\vec{k}\vec{R}}$$

. The corresponding wavefunctions ψ hence can be labeled with the two quantum numbers ϵ and $\vec{k}$ and the relation $\vec{k}(\epsilon)$ for which solutions are found, describes the band structure.

Similar to the unbound spectrum of energy eigenvalues for a give $\vec{k}$ value as usually discussed in band structure theory, there is also an infinite number of wavevectors $\vec{k}$ for a given energy ϵ. However, most of them are not real but complex and hence correspond to decaying or evanescent states. In periodic solids such wavefunction can not occur as they are not normalisable, but in

situations, in which the periodicity is broken at some interface these solutions become relevant and the matching scheme that we employed for the simple examples of barriers before have to take such solutions into account. Hence, it is not the simple relation of decay as a function of barrier height and effective mass that determines tunneling through real materials, but the decay of the wavefunctions is described by the complex band structure that summarizes the solutions $\epsilon(\vec{k})$ with complex $\vec{k}$.
An example of such a complex band structure for an oxide is given in Fig. 8. While we will not discuss all features found in such a plot, we should at least point out a couple of key features that are characteristic for complex band structures:

- In the same way as the usual Bloch bands, the bands with complex $\vec{k}$ (complex bands/branches) are continous.
- The real and complex branches of the bands meet at extrema of the band structure. Hence, high symmetry points are usually of significance.
- The complex branches often form loops across gaps in the real bands. This is of significance for tunneling, as the tunnel barriers correspond to bandgaps in the Bloch bands.
- At very high and low energies one rediscovers the free electron behaviour as the potential become irrelevant, i.e. one finds $\epsilon = \frac{\hbar^2}{2m}\vec{k}^2$.

Many more details on the properties of complex band structures can be found in the literature[20].

7 Applications

Before closing this discussion we will shortly discuss the relevance of the tunneling effect for modern device concepts and give two very basic examples. A key idea of novel devices in nanoelectronics is to change the resistance of device due to some change of configuration. Most of the contributions to this years Spring school deal with such changes due to rather drastic changes of the atomic configuration of the device and in quite some of those the tunneling of electrons can be an important effect. However, there are two concepts in which rather minor modifications of an otherwise perfect tunneljunction is sufficient to have a substantial change in resistivity. These are the tunneling magneto-resistance (TMR) effect and the tunneling electroresistance (TER) effect. The TMR deals with the change in resistivity due to the change of the magnetic configuration, while the TER effect describes the situation in which a ferroelectric barrier is modified.

7.1 Tunneling magneto-resistance

In a magnetic tunneljunction (MTJ) the two metallic leads are magnetic materials and one considers the change in resistance between a parallel alignment of the two magnetisations and an antiparallel alignment. The TMR ratio is then defined as the normalized ratio of the resistance between these two states

$$TMR = \frac{R_{AP} - RP}{R_{AP} + R_P},$$

or, if you are looking for more spectacular numbers by

$$TMR = \frac{R_{AP} - RP}{min(R_{AP}, R_P)}.$$

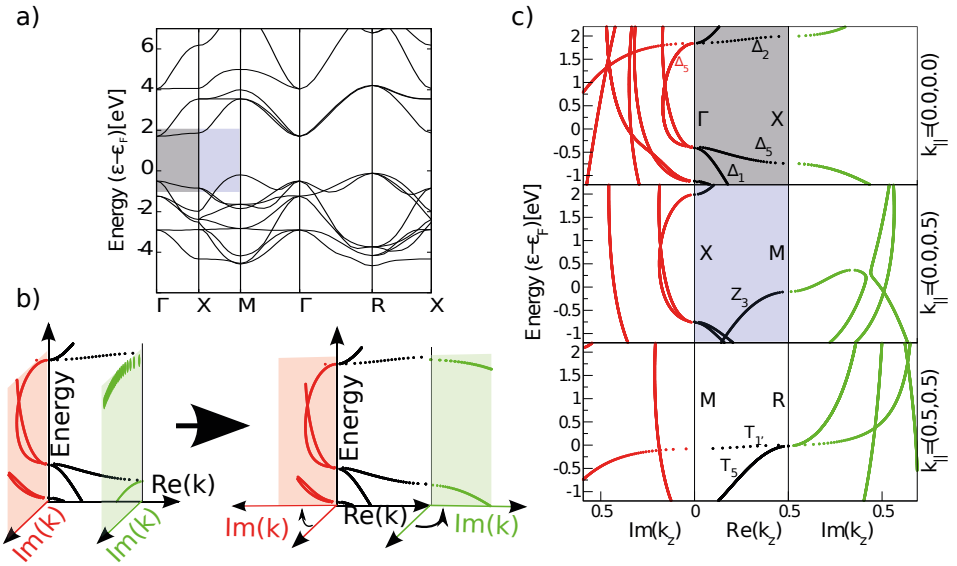


Fig. 8: *Complex band structure of $SrTiO_3$. a) Usual band structure (BS) of the transition metal oxide in the cubic perovskite structure. b) Schematics of the construction of the plot of the complex BS. Additional bands with complex* **k***-values appear at right angle in complex* **k***-space at the high-symmetry points. These bands in the red and green panel are then plotted adjacent to the Bloch states by rotating these $Im(k), \epsilon$ panels next to the $Re(k), \epsilon$ plot. c) complex BS for some special lines in* **k***-space. The parts of the real Bloch BS which is also shown in a) is indicated by shaded areas.*

In such a MTJ the additional spin-degree of freedom of the electrons has to be included in the description of the transport process. In a ferromagnetic metal electrons of different spin exhibit a different electronic band structure, their wavefunctions and transport properties differ. As the electronic density of states for the two spins differ, usually also the number of states relevant for transport, i.e. those states at or close to the Fermi level differ for the two spins. In the problem we are interested in here, the electronic tunneling through an insulating barrier, one usually assumes that the spin of the electrons is not altered. Hence, we assume that the electron spin is a conserved quantum number, no scattering processes coupling electrons of different spins are included in our theory. While this approximation is justified in most cases one should be aware of its limitations. As the tunneling process itself is a coupling phenomena with a very low transition rate, the neglected spin-flip scattering can actually become a major effect as it was demonstrated e.g. in the case of surface states in half-metallic MTJs.
If the spin is not changed during the transport process across the MTJ, one can decompose the total current into a spin-up and a spin-down component

$$I = I_\uparrow + I_\downarrow \tag{40}$$

in which the up-spin electrons form $I_\uparrow$ and the down-spin electrons $I_\downarrow$. This two-current model, which can be equivalently expressed as a "two resistor" model in which the junctions is consid-

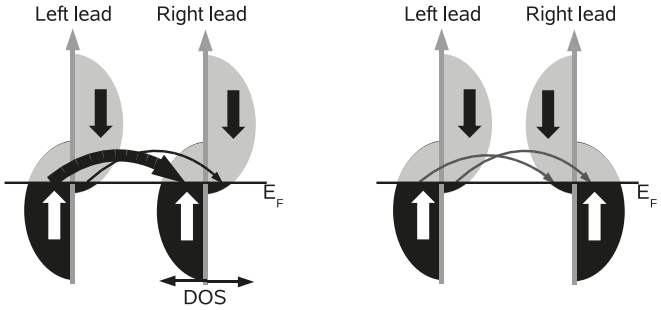


Fig. 9: *Julliere model of spin-polarized tunneling. The spin-polarized density of states(DOS) of the parallel(left) and antiparallel(right) MTJ is sketched. The tunneling current is indicated by curved arrows at the Fermi energy E_F. In the parallel situation the large DOS at both sides of the junction in the ↑-spin leads to a large current while all other currents are small due to the small number of available states.*

ered as two parallel resistors with $\frac{1}{R} = \frac{1}{R_\uparrow} + \frac{1}{R_\downarrow}$, is an extremely popular and successful concept in spintronics. Starting from this ansatz Julliere [19] constructed a very basic model explaining the TMR effect. His basic assumption was that the current across the junction is proportional to the product of some density of states(DOS) of the two sides (see Fig. 9), i.e. including the two current model one obtains the following expression for the total current

$$I_P \propto n_{L\uparrow} n_{R\uparrow} + n_{L\downarrow} n_{R\downarrow}. \tag{41}$$

This we will assume to be the expression for the parallel current, if we now switch to an antiparallel alignment to of the electrodes, we will have to flip the spins of one side with respect to the other. If we assign this switching to the right electrode we would obtain the following expression for the current in the antiparallel case

$$I_{AP} \propto n_{L\uparrow} n_{R\downarrow} + n_{L\downarrow} n_{R\uparrow} \tag{42}$$

and consequently we get for the TMR value

$$TMR = \frac{I_P - I_{AP}}{I_P + I_{AP}} = \frac{(n_{L\uparrow} - n_{L\downarrow})(n_{R\uparrow} - n_{R\downarrow})}{(n_{L\uparrow} + n_{L\downarrow})(n_{R\uparrow} + n_{R\downarrow})} = P_L P_R, \tag{43}$$

where P_L and P_R are the spin-polarizations of the density of states

$$P_{L/R} = \frac{(n_{L/R\uparrow} - n_{L/R\downarrow})}{(n_{L/R\uparrow} + n_{L/R\downarrow})} \tag{44}$$

for the left and right electrodes, respectively. This very simple expression, known as Julliere's formula can already explain many basic features of TMR:

- TMR only occurs if both electrodes are magnetic, i.e. if both have a non-vanishing spin-polarization.
- The maximal TMR of 100% is expected if both electrodes are 100% spin-polarized, i.e. if both electrodes behave like half-metals.

- If both electrodes are equivalent, i.e. if $P_L = P_R$, the TMR effect is always positive.

At the same time one can easily spot several shortcomings of this theory of which the most significant is the use of the rather ill-defined densities of states $n_{L/R}$. Obviously, these quantities have to be somehow related to the electronic density of states at or close to the Fermi level as these electrons will carry the electric current. Furthermore, it must be somehow related to the local density of states at the metal/insulator interface since this is the region of space from which tunneling "takes place". However, the exact definition of these quantities is unclear and in consequence the predictive and explanatory power of Julliere's formula is strongly limited.

7.2 Tunneling electro-resistance

Similar to the MTJ, in which the change of the direction of the magnetization will modify the resistance of the junction, one can obtain a similar effect if one considers a tunneljunction including a ferroelectric that can change its direction of polarization. As ferroelectrics are insulators, they will of course not be the materials for the leads but for the tunnelbarrier itself. In the case of a ferroelectric barrier material, the electric polarization within the barrier can lead to a non-constant barrier height. Neglecting all local variations of the polarization within the barrier one would obtain a barrier potential that changes linearly within the barrier with some finite slope. Hence, one aims at generalizing the result of we obtained for the rectangular barrier by simply assuming that the decay constant κ now varies within the barrier. The simplest generalization, which can also be interpreted at the well-known Wenzel-Kramers-Brillouin (WKB) approximation for the tunneling of electrons, is given by

$$\Gamma = C\, e^{-2\int_0^d \kappa(z)dz} = C\, e^{-2\overline{\kappa}d}, \tag{45}$$

where $\overline{\kappa}$ is simply the averaged decay constant across the barrier and the constant C contains all interface details.
When switching the polarization direction, the barrier potential will change. Due to time-inversion symmetry, this can only lead to a change in tunneling conductance if the total tunnel junction is not symmetric, i.e. if the barrier after switching the polarization is not simply the mirror image of the unswitched barrier. This requires that the two interfaces between the leads and the ferroelectric barrier are different, so that the induced screening charges within the metallic leads and/or the details of the chemical and atomic arrangement is different[18]. In the asymmetric case, the shape of the potential barrier will be different for the two directions of the polarization leading to a tunneling electro resistance given by

$$TER = \frac{\Gamma_{\leftarrow} - \Gamma_{\rightarrow}}{\Gamma_{\leftarrow} + \Gamma_{\rightarrow}} \approx \tanh d\,(\overline{\kappa_{\rightarrow}} - \overline{\kappa_{\leftarrow}}) \tag{46}$$

(here expressed by the relative change in the conductance).
Here we again assumed that the matching at the interface is the same for both polarization directions, i.e. the factor C in Eq. 45 is the same for the conductance in both directions of polarization $\Gamma_{\leftarrow}$ and $\Gamma_{\rightarrow}$. In the effective mass model one can easily evaluate the average decay constant and the resulting TER. Assuming that the barrier height is changing linearly across the barrier, the decay constant is given by $\kappa(z) = \frac{1}{\hbar}\sqrt{\frac{2m}{d}(V_1(d-z) + V_2 z)}$, where V_1 and V_2 are the barrier heights at both ends of the tunneling barrier (i.e. at z=0 and z=d)(see Fig. 10). The

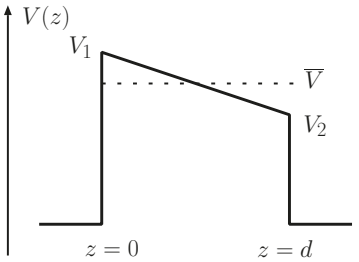


Fig. 10: *Simplest model of a potential barrier with a ferroelectric barrier. The barrier height is measured with respect to the energy of the propagating electrons in the leads.*

resulting averaged decay constant is then

$$\overline{\kappa} = \frac{2}{3}\frac{1}{\hbar}\frac{\sqrt{2m}(V_1^{3/2} - V_2^{3/2})}{(V_1 - V_2)}. \tag{47}$$

and the tunneling electro resistance ratio can be expressed in terms of the barrier heights. Introducing the averaged barrier height $\overline{V}$ and setting $V_1 = \overline{V} + \frac{1}{2}\Delta V$ and $V_2 = \overline{V} - \frac{1}{2}\Delta V$ one obtains up to order of ΔV^4

$$\overline{\kappa} \approx \frac{1}{\hbar}\sqrt{2m\overline{V}} - \frac{1}{\hbar}\sqrt{2m}\frac{1}{6\overline{V}^{3/2}}\Delta V^2. \tag{48}$$

This simple analysis demonstrates a few basic features of the TER due to the change of the barrier potential:

- A change in the average potential height $\overline{V}$ will lead to a variation of the tunneling transmission. While the effect of this change of average potential can be expected to be strong, it will usually be accompanied by a change in energy of the system, i.e. the system will have a preferred orientation of the polarization which might not be a desirable effect for many application scenarios.

- A change of slope of the barrier, i.e. a modification of ΔV, will result in a TER effect. This is due to the non-linearity of the decay constant as a function of barrier height and the effect will depend on the average barrier height. A variation close to the band edge, i.e. with a small average potential $\overline{V}$ will be more significant as one can expect from the parabolic dispersion of the decaying band.

- As the TER is due to difference of the electronic structure in the barrier, it scales with the barrier thickness (see Eq. 46). This is a fundamental difference to the TMR in which the spin-polarization of the leads and the interface are fundamental for the effect. One should also note here, that the TER does not directly depend on a residual field in the barrier. Such a field will decay with increasing thickness. As the TER depends on the potential alignment at the interfaces it can be expected to be largely independent of the barrier thickness.

While this simple model indicates some basic features of TER, we have to stress, that it is quite an inappropriate description of realistic devices. For example a simple junction, in which the barrier is composed of $BaTiO_3$ the difference between the decay constants calculated from the simple model here and the values obtained from a realistic complex band structure is drastic. As one can see from Fig. 11, the complex bands with smallest decay constant, i.e. those most relevant for tunneling have a very small energy dependence. These nearly vertical bands lead to a decay across the barrier that is roughly independent of the energy within the bandgap. Hence, the arguments in our simple model which relied on the change of decay with different barrier heights do not apply and one would obtain no TER in such a junction. On the other hand however, the differences in the interface details upon switching of the polarization of course could still lead to a TER, but such effects would have to be investigated for each particular setup and no general conclusions can be easily drawn here.

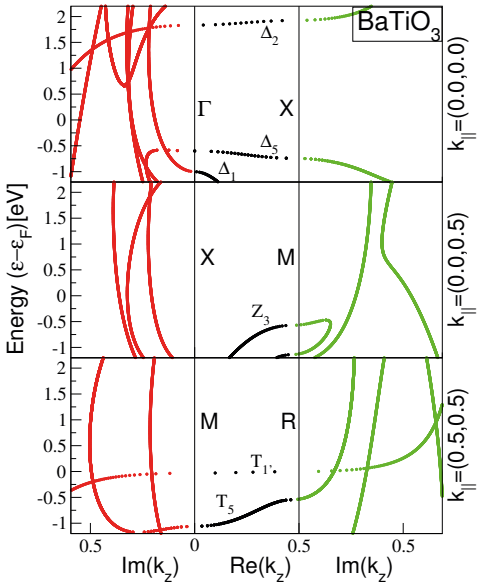


Fig. 11: *Complex band structure of $BaTiO_3$. As the complex bands with smallest decay in the bandgap are not directly derived from the bottom of the conduction band, the decay is only weakly dependent on energy, i. the bands are nearly vertically in the energy range of the bandgap.*

References

[1] L. Nordheim. Zur Theorie der thermischen Emission und der Reflektion von Elektronen an Metallen. Z. Phys. 46:833 (1927).

[2] G. Gamow. Zur Quantentheorie des Atomkernes. Z. Phys. 51:204 (1928).

[3] R. W. Gurney and E. U. Condon. Quantum machanics and radioactive disintegration. Phys. Rev. 33:127 (1929).

[4] O. Wunnicke, N. Papanikolaou, R. Zeller, P. H. Dederichs, V. Drchal and J. Kudrnovský. Effects of resonant interface states on tunneling magnetoresistance. Phys. Rev. B 65:064425 (2002).

[5] R. Landauer. Spatial Variation of Currents and Fields Due to Localized Scatterers in Metallic Conduction. *IBM Journal Res. Dev.*, 1:223, 1957.

[6] M. Büttiker. Four-terminal phase coherent conductance. *Phys. Rev. Lett.*, 57:1761, 1986.

[7] A.D. Stone and A. Szafer. What is measured when you measure a resistance? - The landauer formula revisited. *IBM Journal Res. Dev.*, 32:384, 1988.

[8] S. Heinze, S. Blügel, R. Pascal, M. Bode, and R. Wiesendanger. Prediction of bias-voltage-dependent corrugation reversal for STM images of bcc (110) surfaces: W(110), Ta(110), and Fe(110). *Phys. Rev. B*, 58:16432, 1998.

[9] S. Heinze, R. Abt, S. Blügel, G. Gilarowski, and H. Niehus. Scanning tunneling microscopy images of transition-metal structures buried below nobel-metal surfaces. *Phys. Rev. Lett.*, 83:4808, 1999.

[10] B. Voigtländer, V. Scheuch, H.P. Bonzel, S. Heinze, and S. Blügel. Chemical identification of atoms at multicomponent surfaces on an atomic scale: $CoSi_2$(100). *Phys. Rev. B*, 55:R13444, 1997.

[11] V. P. LaBella, H. Yang, D. W. Bullock, P. M. Thibabo, Peter Kratzer, and Matthias Scheffler. Atomic structure of the GaAs(001)-(2×4) surface resolved using scanning tunneling microscopy and first-principles theory. *Phys. Rev. Lett.*, 83:2989, 1999.

[12] F. Schwabl. *Quantenmechanik*. Springer, 4 edition, 1993.

[13] D. Wortmann, H. Ishida, and S. Blügel. *Phys. Rev. B*, 66:075113, 2002.

[14] D. Wortmann, H. Ishida, and S. Blügel. *Phys. Rev. B*, 65:165103, 2002.

[15] http://www.flapw.de.

[16] A. Eiguren, B. Hellsing, E. V. Chulkov, and P. M. Echenique. Phonon-mediated decay of metal surface states. *Phys. Rev. B*, 67:235423, 2003.

[17] S. Link, H. A. Dürr, G. Bihlmayer, S. Blügel, W. Eberhardt, E. V. Chulkov, V. M. Silkin, and P. M. Echenique. Femtosecond electron dynamics of image-potential states on clean and oxygen-covered Pt(111). *Phys. Rev. B*, 63:115420, 2001.

[18] E. Y. Tsymbal and H. Kohlstedt. Tunneling across a ferroelectric *Science*, 313, 181 (2006), ISSN 1095-9203.

[19] M. Julliere. *Phys. Lett. A* 54A:225 (1975).

[20] V. Heine. *Surface Science* 2:1 (1964).

B 1 Chemical Vapour Deposition Techniques

Susanne Hoffmann-Eifert
Peter Grünberg Institute, PGI 7
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Concept and tools **4**
2.1 Precursor chemistry 4
2.2 The precursor supply system 7
2.3 The reactor design 8

3 Metal organic chemical vapour deposition (MOCVD) **9**
3.1 MOCVD principle 9
3.2 Growth rate limitation 10
3.3 Gas flow schemes 11

4 Atomic layer deposition (ALD) **12**
4.1 Self-limiting surface reactions 12
4.2 ALD (temperature) window 13
4.3 Uniformity and conformality 14
4.4 Advanced ALD process schemes 15
4.5 ALD co-reactants 15

5 ALD chemical reactions **16**
5.1 Al_2O_3 ALD 17
5.2 HfO_2 ALD 18
5.3 GeTe ALD 19

6 Growth control **20**
6.1 Uniformity and conformality 20
6.2 Early stages of growth – towards ultrathin films 21
6.3 Growth at low temperatures 22
6.4 Microstructure and phase control 22

7 Summary **23**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

Memristive devices are in general built form artificial layer stacks of highest integration density comprising metal/resistive switching layer/metal structures. The memristive principle of information storage is not related to the amount of stored charge like for example in dynamic random access memory (DRAM) or field effect transistors (FET) but it is based on a change in the conductivity of an individual cell representing one bit.[1] Therefore, in contrast to DRAM technology, the device area of memristive devices can be reduced at least to 100 nm^2 with a thickness of about 3 to 10 nm for the functional layer, and about 30 nm for the complete stack (see Fig. 1).[2] Highest integration density of $4F^2$, where F is the feature size, is obtained in three dimensional integrated passive crossbar arrays as shown in Fig. 2.[3] The nanometer-size dimensions make the deposition of the thin film functional layers an essential step in memristive devices fabrication.

When considering polycrystalline and amorphous films for microelectronic applications, the use of vapour phase deposition techniques such as chemical vapour deposition (CVD) and physical vapour deposition (PVD) is a prior condition. PVD (to be discussed in Chapter B2) describes a variety of techniques that are based on the condensation of a vaporized form of the film material on the substrate. This vaporized material is obtained from a target by purely physical processes, such as thermal or laser-induced evaporation or sputtering by energetic ion bombardment. On the other hand, CVD methods (the topic of this paper) involve chemical reactions. These chemical reactions take place by volatile precursor molecules that decompose at the surface leaving behind a thin film and volatile by-products. The chemical reactions are thermally driven, most frequently by heating the substrate, and can also be enhanced by reactive species created in the gas phase, e.g. in plasma-enhanced CVD. It should be mentioned that additional integration techniques, like for example lithography methods, structuring and etching are equally important for the fabrication of the nanometer size structures but are beyond the scope of this chapter.

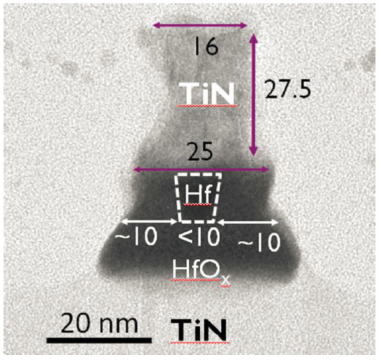


Fig. 1: *Transmission electron micrograph showing a TiN/HfO_x/Hf/TiN resistive element with a width of less than 10 nm. [2]*

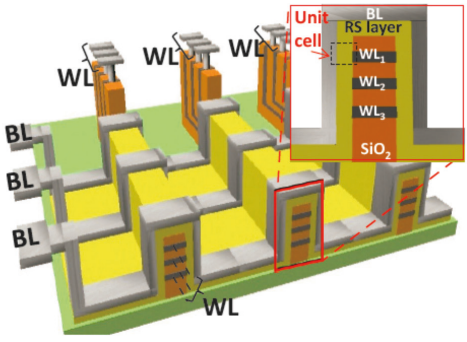


Fig. 2: *Schematic of a 3D vertical RRAM array. The unit cell can be realized using a Ta/TaO_x/TiO_2/Ti cell. [3]*

The continuing trend in the miniaturization of the critical device dimensions and the processing of devices on increasingly larger substrates sets stricter and new demands on the film deposition methods which leads to a requirement for a growth control in terms of three metrics, illustrated in Fig. 3.[9]

- *Thickness control*: The deposition of high-quality ultra-thin films with a thickness control at the sub-nanometer level.
- *Uniformity and conformality*: The uniformity of the films on large wafers and good conformality for surface features including trenches, pores, surface roughness, etc.
- *Low temperature*: The ability to deposit high-quality materials with a high purity and a high density (no voids or pinholes) at low substrate temperatures.

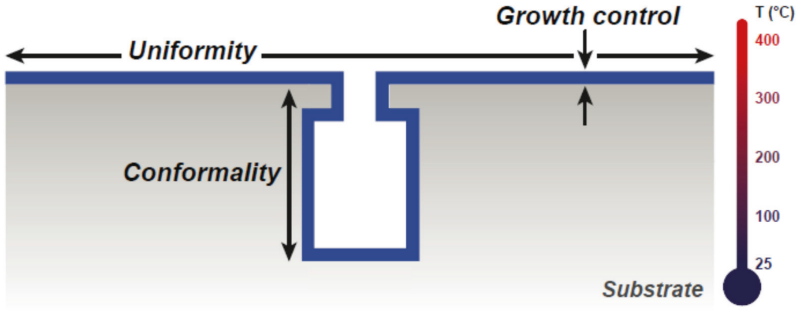


Fig. 3: *The coverage metrics of a thin film on a substrate with three-dimensional (3D) features. Uniformity/conformality give the coverage of the planar surface/3D features, respectively. The growth control over the film thickness itself is important and the ability to achieve these metrics at low temperatures.(taken from [9])*

In this chapter we will discuss basic principles of chemical vapour deposition methods which are utilized for the growth of *polycrystalline* and *amorphous* metal oxide (MO), transition metal oxide (TMO) and phase change (PCM) layers, namely metal organic chemical vapour deposition (MOCVD) and atomic layer deposition (ALD). First, general aspects of chemical vapour deposition will be introduced, different types of evaporation systems, reactor concepts, and deposition modes will be shown and most relevant chemical precursors will be summarized. Differences between MOCVD and ALD will be outlined, and some details of the techniques will be discussed. At the end, several examples of integrated resistive switching films obtained from ALD will be given.

In the frame of this lecture only fractions of the fields can be touched. For further detailed study the reader is kindly referred to the excellent books like '*Chemical Vapour Deposition: Precursors, Processes and Applications*' edited by Jones & Hitchman [4], '*Atomic Layer Deposition of Nanostructured Materials*' edited by Pinna & Knez [5], '*Atomic Layer Deposition for Semiconductors*' edited by Hwang [6], and review articles from leading groups in the field with focus on *precursor chemistry* by Devi [7], on *ALD processes* by Miikkulainen et al. [8], and on different *ALD techniques* by Knoops et al. [9].

2 Concept and tools

Chemical vapour deposition (CVD) and atomic layer deposition (ALD) have emerged as effective techniques for the growth of thin films offering several advantages in terms of film quality and scale up. The concept of CVD can be described by a transformation of molecules to materials, i.e. inorganic thin films which grow on a substrate material. The molecules are transferred into the gas phase, and the precursor vapour is transported to the substrate by means of a gas stream or only by an established pressure gradient. The chemistry of the starting molecules plays a vital role in the deposition process by governing the chemical reactions which control the formation of the inorganic film. While the differences of CVD and ALD principles are discussed later, this paragraph will give a short introduction to common precursors, methods of precursor supply and reactor concepts.

2.1 Precursor chemistry

A common requirement for all precursors utilized in CVD processes is a *high vapour pressure* at a moderate temperature which is below the sufficiently high decomposition temperature of the molecules. The range between both temperatures defines the *process window*. The choice of a suitable chemical precursor is governed by certain general characteristics which are summarized in Fig. 4.[7] Metalorganic chemical vapour deposition (MOCVD) is one variant of CVD where metalorganic precursors are used. The most important properties that any CVD precursor has to possess are adequate *volatility* and a sufficiently large temperature "window" between evaporation and decomposition for film deposition. This is usually not a problem for gaseous precursors but might become an issue for liquid and solid precursors. Volatility of CVD precursors can often be enhanced by introducing bulky ligands, which reduce the inter-molecular forces that lead to dimer, oligomer or polymer formation.

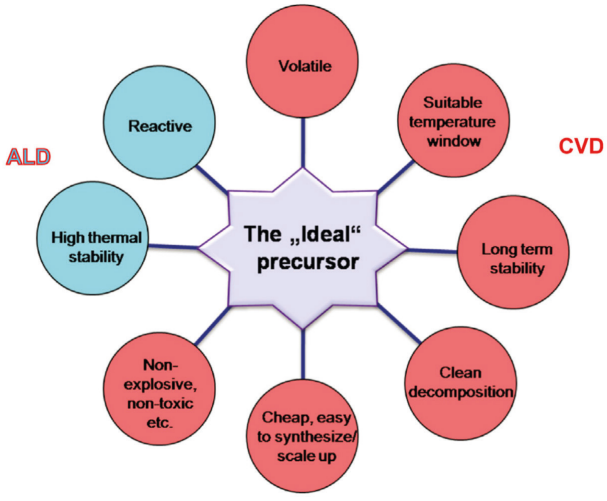


Fig. 4: *Summary of the most important characteristics for an ideal CVD (red) and ALD (red + blue) precursor. (taken from [7])*

Another way to overcome this problem is to use the liquid injection CVD technique, where the given compound is dissolved in a suitable solvent, usually toluene, ethylcyclohexane, THF, n-butylacetate etc. Evaporation of the compound is achieved by flash evaporation of the precursor solution. In addition, a CVD precursor needs to be sufficiently stable at room temperature when stored over a long period of time, and it should not undergo any decomposition at the evaporation temperatures necessary to achieve adequate gas-phase transport of the vapour. Further requirements include a clean decomposition on pyrolysis (see Fig. 6) in order to give desired material with minimum contamination (low carbon content in the case of oxides for micro- and optoelectronic applications) and an easy and cost efficient synthesis based on inexpensive, readily available chemicals. In an ALD process, additionally, sufficient precursor thermal stability is needed, both in the gas phase and on the substrate surface in order to avoid uncontrolled thermal decomposition reactions. Furthermore, the precursor must be reactive towards the surface groups and leave reactive surface groups, and precursor must not react with itself or with its surface-adsorbed species. In that way, it is possible to reach the saturation stage in a short time (within few seconds) and thereby ensure a reasonable deposition rate. To achieve this, the desired ALD reactions should have large negative ΔG values. In general, the use of bidentate ligands usually increases the thermal stability of the precursors, while the stronger Brønsted basic character of the ligand increases precursor reactivity towards water and OH–functionalities.[7] Current metal sources are almost exclusively inorganic coordination complexes, i.e., a metal centre surrounded by ligands. Examples of typical inorganic complexes used as precursors are shown in Fig. 5.[9] The ligands are essentially what make the metal centre volatile and they play a key role in determining the characteristics of the precursor.

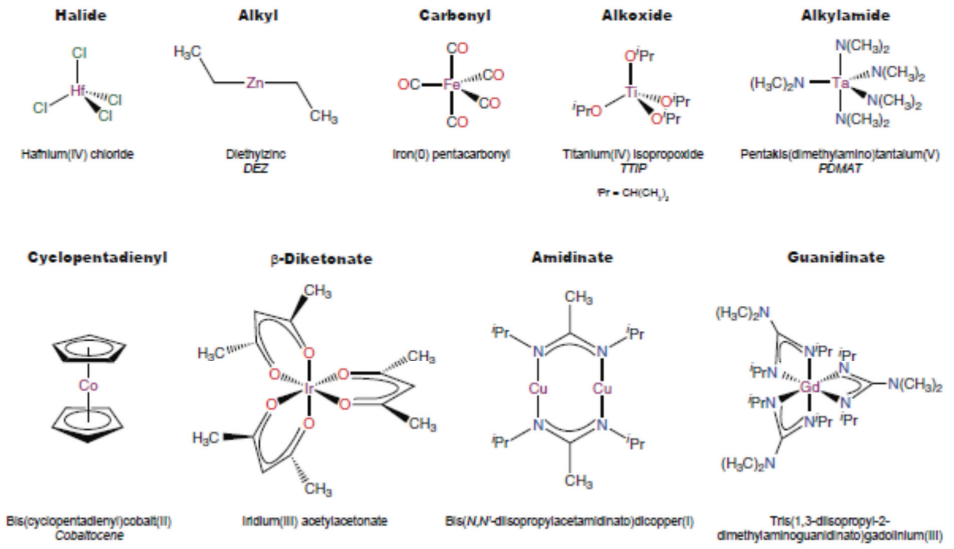


Fig. 5: *Examples of compounds used as the metal source in MOCVD and ALD. The names in bold refer to the general class of precursor compound. (taken from [9])*

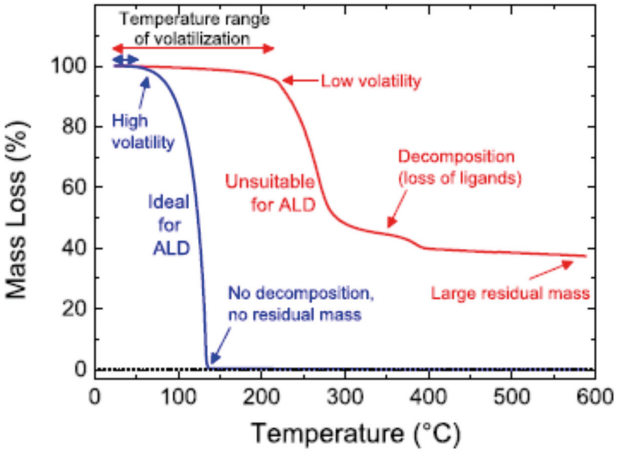


Fig. 6: *Thermogravimetric analysis (a compound's mass loss as a function of temperature at a set heating rate, usually 10 °C/min) of an ideal (blue line) and a non-ideal (red line) precursor. (taken from [9])*

Some of the aforementioned precursor considerations with respect to volatility are outlined in the thermogravimetric analysis (TG) curves in Fig. 6, which show the compounds' weight losses as a function of temperature. Ideal precursors should be volatile at low temperatures and must not decompose easily, which is characterized by a swift mass drop to 0% (blue curve in Fig. 6). However, there are many compounds that exhibit incomplete decomposition (red curve), where the volatilization takes place slowly over a wide temperature range. In this case the mass loss is likely a sign of decomposition, observable by a large residual mass and loss of ligands, denoted by multiple plateaus and mass drops in the TG curve. Such compounds are unsuitable for CVD application. As a general (not exclusive) rule, heavier, symmetrical molecules tend to exhibit low volatility, whereas lighter, highly asymmetric compounds are more volatile. A compound's volatility is affected by intermolecular forces, such as hydrogen-bonding or electrostatic interactions (mainly van der Waals forces), which in turn are influenced in varying degrees by the molecular weight of the precursor and the shape of the molecule. A standard method of improving volatility is to adopt a *heteroleptic* precursor (i.e., a metal centre with two or more different ligands), which introduces asymmetry, as opposed to a *homoleptic* molecule (where all the ligands are the same).

Halide-based precursors, such as $HfCl_4$ (see Fig. 5), are highly desirable for industrial processes as they are highly reactive, stable compounds and are relatively inexpensive. However, halide ligands have a high tendency to contaminate films, especially at low deposition temperatures. Additionally, the reaction products, like e.g. HCl, can be corrosive toward both the film and the reactor. Organometallic compounds, such as metal alkyls (e.g. $\{Al(CH_3)_3\}_2$, trimethyl aluminium, TMA) or metal carbonyls (e.g., $Fe(CO)_5$), do not have these problems and are still highly reactive with high vapour pressures. However, the low energy of the M–C bond affects their thermal stability and shelf life, potentially making handling difficult; for example, metal alkyls tend to be pyrophoric, and metal carbonyls readily decompose, even at room temperature. A stronger M–C bond is obtained for metal cyclopentadienyl (Cp) complexes (e.g., $FeCp_2$), as the Cp ligand helps to saturate the metal centre both electronically and coordinatively through π-bonding. Other popular catego-

ries of precursor include alkoxides (e.g. $Ti(OiPr)_4$, titanium tetra isopropoxide, TTIP) and alkylamides (e.g. $Ta(N(CH_3)_2)_5$, pentakis (dimethylamino)tantalum, PDMAT). Both types of precursors can oligomerize. In order to overcome this, ligands that bind to the metal centre via two or more atoms can be adopted to block vacant coordination sites on the metal. This can either be to the same metal centre (a chelating ligand), as exhibited by β-diketonates (and their analogues, β-diketiminates and β-ketoiminates), amidinates, and guanidinates, or by a ligand bridging across two metal centres. More details on precursor chemistry can be found in [7] and [8].

2.2 The precursor supply system

The precursor supply system has to be customized with respect to the volatility of the selected type of molecule. Different techniques shown in Fig. 7 have been developed to vaporize the precursors and to transfer them into the reaction chamber (1) direct delivery, (2) bubbling, and (3) liquid injection. For the methods (1) and (2) the vapour pressure can be increased by moderate heating. Precursors with reasonable vapour pressure in the process window are directly delivered to the deposition chamber. A slightly low vapour pressure can be increased by moderate heating or/and bubbling of the precursor. In the latter case inert gas flows through the precursor which causes a reduction of the surface tension and in consequence an easier transfer of molecules into the gas phase which results in an increase in vapour pressure at a given temperature. Solid precursors are sublimed. However, an increase of the vapour pressure by means of the bubbling technique is difficult, although concepts are available on the market. If, on the other hand, the precursor shows a good solubility in an inert solvent like for example toluene, the dissolved precursor can be delivered by means of the liquid injection technique (LI) using for example a spray injection system. Via a nozzle the precursor solution is sprayed into a heated tube where the small droplets are transported by means of a carrier gas. Passing the heated tube, every droplet is smoothly vaporized and the precursor molecules are transferred into the gas phase. The precursor gas is then transported to the reaction chamber where the MOCVD or ALD reactions take place.

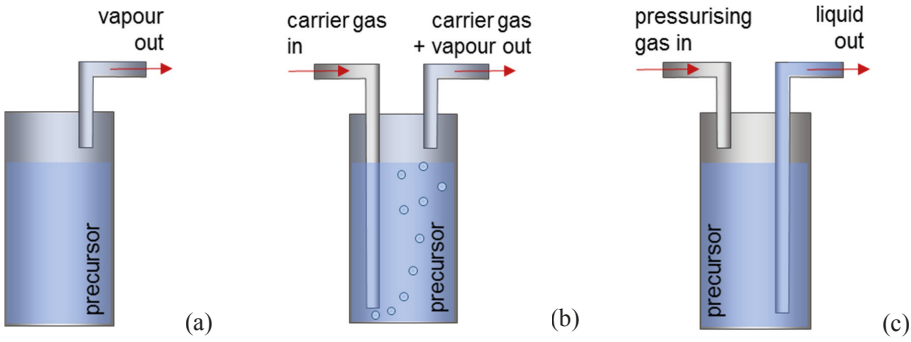


Fig. 7: *Precusor delivery modes a) direct delivery of the precursor gas, b) bubbler-type precursor supply, c) liquid injection mode where the solution is delivered to the vaporizer. For the methods (1) and (2) the vapour pressure can be increased by moderate heating.*

2.3 The reactor design

In general, the reactor design has to support a continuous, completely homogenous, and highly reproducible deposition reaction. Therefore, the reactor must provide a controlled gas flow and heat distribution and, for the precursors discussed here, it should work at low pressure (10^{-1} to 10 mbar).[4] For small wafers, a horizontal gas flow is most appropriate which can be realized in a linear reactor for single wafers or in a multi-wafer planetary reactor.[10] Another system especially suitable for very large (e.g. 300 mm) wafers is the shower-head design (see Fig. 8). Here, the precursor vapour is distributed over a temperature controlled plate with numerous defined holes (so-called ´showerhead´) which supplies the precursor vapour homogeneously over a large area.[9] The CVD reactor design and geometry has a very high influence on the film growth regarding homogeneity and reproducibility. It is designed by numerical solution of gas fluid dynamics and reaction processes. For a precise control of the material composition, especially for multicomponent materials, the vapour flows have to be exactly controlled in addition to the substrate temperature (see Fig. 9). The flow rate of precursor, f_p, given in sccm (standard cm^3 per minute) is:

$$f_p = \frac{f_c \, p_p}{p_{tot} - p_c} \tag{1}$$

The index c refers to the carrier gas. The partial pressure of the precursor, p_p, can be calculated from the enthalpy of evaporation, ΔH, by means of the Clausius-Clapeyron equation:

$$\frac{d(\ln p)}{dT} = \frac{\Delta H}{k_B T} \tag{2}$$

For stable temperatures, the flow of precursor can be controlled by the flow of the carrier gas.

In contrast, ALD is controlled by chemisorption of the precursor molecules on the substrate surface. Excess of physisorbed molecules is purged away by a gas stream. Due to this saturation-type deposition scheme, for a pure ALD reactor it is preferable, but not necessary, to have a precisely homogeneous gas flow if excess material is purged away prior to the next half reaction. Highest priority for ALD reactors is a small reactor volume which enables fast purging and pumping.

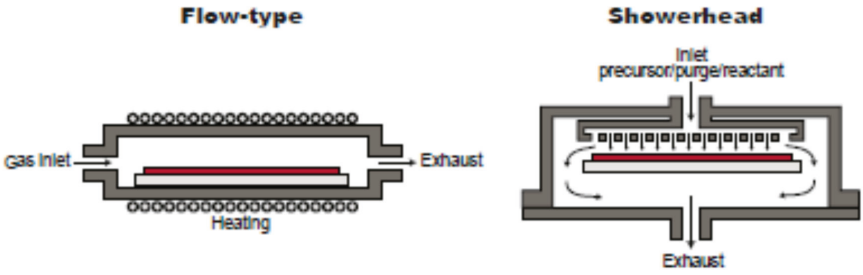


Fig. 8: *Schematic of single wafer reactors: flow type (left) and showerhead reactor (right). [9]*

3 Metal organic chemical vapour deposition (MOCVD)

In MOCVD, the film growth occurs through the chemical reaction of precursor molecules which are transported to the vicinity of the substrate via the vapour phase. The film building chemical reactions typically utilize thermal energy from the heated substrate (see Fig. 9). For growth processes under a limited thermal budget, a part of the necessary energy to drive the reaction can be supplied from non-thermal energy sources such as radio frequency (RF) or microwave power or (UV) light. The advantages of MOCVD are (1) the opportunity to deposit epitaxial thin films relatively easily because of the molecular reaction, lay-down, and incorporation into the crystal lattice at the surface, (2) the homogeneous deposition over large areas for some of the reactor designs, (3) the compatibility with the semiconductor-fabrication techniques, and (4) the opportunity to achieve a reasonable step coverage even for 3D structures of moderate high aspect ratio.

3.1 MOCVD principle

MOCVD of oxide thin films is based on the evaporation of the precursors which are decomposing close to the hot substrate in a suitable reaction chamber in which the temperatures, the pressure of gases such as oxygen or other oxidizing gas, inert gases, and the precursor vapor are controlled (see Fig. 9). A complete understanding of the complex MOCVD process includes processes on different length scales [10]: on the macroscale (m) finite element methods are used to calculate the continuum fluid flow, heat and mass transfer; on the microscale (µm) Monte Carlo methods provide calculations of surface diffusion and kinetics while on the atomic scale (nm) microscopic chemical dynamics is considered. The deposition rate and the final composition of the film are primarily controlled by the spatial profiles of gas velocities, temperatures, and partial pressures of the various precursors in the reactor. Collectively, these profiles are referred to as the reactor flow pattern. For MOCVD-type processes a laminar gas flow pattern is required.[4]

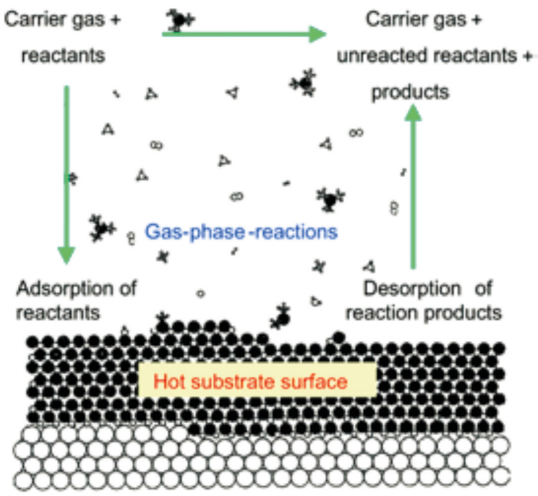


Fig. 9: *Schematics of the gas flow and the atomic scale chemical environment in the region of the growing film surface during an MOCVD process. (taken from [10])*

3.2 Growth rate limitation

Film growth in MOCVD is primarily fueled by thermal energy provided by the heated substrate. Depending on the temperature, two different regimes for growth rate limitation can be identified as shown in Fig. 10. At low temperatures, the growth is usually limited by the *reaction kinetics* (line A) and at higher temperatures by the *mass transport* (line B).

In general, the kinetically limited growth rate, j_k, can be written as

$$j_k = const_1 \cdot N^{\infty} \cdot \exp\left(-\frac{W_a}{k_B T}\right) \tag{3}$$

The effective activation energy, W_a, summarises different process steps ranging from precursor reactions, e.g. decomposition, to the film growth kinetics. N^{∞} gives the precursor concentration at some distance from the substrate.

At medium and higher temperatures major limitations of the reaction are given by the restricted transport of the precursors and reactants to the surfaces, i.e. by the diffusivity, D, of the vapour. The transport limited growth rate, j_t, is given by

$$j_t = const_2 \cdot N^{\infty} \cdot \frac{\sqrt{D}}{T} \tag{4}$$

The diffusion constant of the gas is $D = \lambda \langle u \rangle / 3$, and $\langle u \rangle$ is the average velocity of the gas atoms. The diffusivity of gases shows only a weak temperature dependence, and consequently also j_t. Calculations yield $j_t \sim T^{-1/6}$.

The overall growth rate, j_g, may be obtained by adding the two reciprocal fluxes:

$$\frac{1}{j_g} = \frac{1}{j_k} + \frac{1}{j_t} \tag{5}$$

From these considerations some general rules can be derived which may guide the selection of the appropriate temperature region.

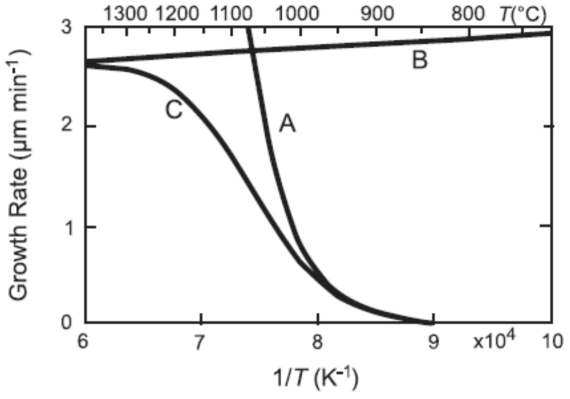


Fig. 10: *Schematic description of the control of the MOCVD growth process; A: kinetic control, B: transport control, C: resulting behaviour.[4]*

Mass transport limited region: The deposition rate j_t is insensitive to small temperature gradients therefore cold wall reactors are often operated in this regime. However, the local deposition rate is very sensitive to the flow pattern and uniformity problems may arise.

Kinetically limited region: The deposition rate j_k is only weakly dependent on the flow homogeneity, therefore the regime is best suited for conformal deposition. However, the exponential temperature dependence necessitates a very high temperature stability and uniformity over large wafers. For MOCVD processes, conformality is expected for kinetically controlled growth, and low temperatures and low pressures are favorable. However, applying this to multicomponent materials results in only a very narrow process window and a high sensitivity to fluctuations in process control. Fig. 11 shows a successful example, i.e. an MOCVD processed $SrTiO_3$ thin layer in an approx. 900 nm deep and 150 nm wide hole in SiO_2 [11]. The bottom to top step coverage is 0.99 and the Sr/Ti stoichiometry variation is within +/- 5%, as determined by EDS TEM.

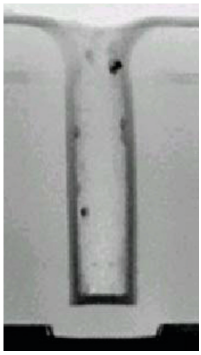


Fig. 11: *MOCVD of $SrTiO_3$ thin films into a test hole of a SiO_2 layer with an aspect ratio of 1:6 and a width of 150 nm. The deposition was performed in a dome-type reactor at a substrate temperature of 420°C. The TEM cross section shows the very high conformity of the thin film.[11]*

3.3 Gas flow schemes

One of the most intuitive differentiation between the various CVD techniques is given by a comparison of the gas flow schemes as depicted in Fig. 12.[12] A typical MOCVD process is characterized by continuous flows of precursor vapour and reactive gas. Whereas, in the special case of liquid injection MOCVD the precursor vapour is supplied to the reaction chamber in a pulse mode which originates from the pulse-type spray evaporation of the precursor solution. This technique is commercialized as 'atomic vapour deposition' (AVD®). Different precursor solutions are often evaporated in sequential pulses. By this gas phase reactions between different metal precursors are nearly avoided and only reactions between a metal precursor and the reactive gas, which is continuously flowing, and the substrate remain. The next step towards complete pulse operation is realized in the ALD technique. Here, both, metal precursor vapour and reactive gas are delivered in a pulse series. Due to the absence of reactants in the gas phase and supported by the low substrate temperature, no gas phase reactions are expected to occur, if reactions between identical precursor molecules and decomposition are avoided. These conditions allow the chemisorption of the precursor with reactive surface sites on the substrate as the only chemical reaction to take place apart from adsorption events. Purge steps between the pulses, precursor vapour or reactive gas, should inhibit any gas phase reactions.

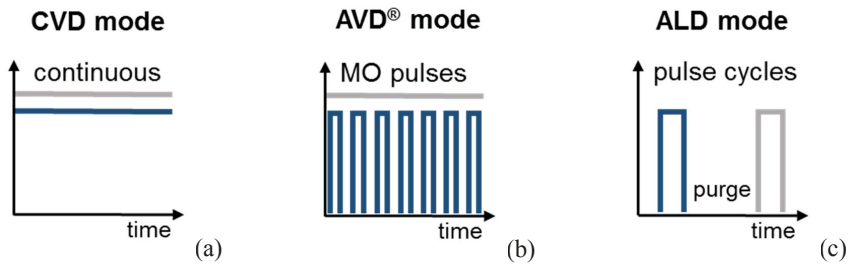


Fig. 12: *Gas flow schemes for typical CVD techniques a) conventional CVD, b) liquid injection MOCVD or AVD®, and c) ALD. The blue and grey lines show the metal precursor vapour and the reactive gas input, respectively. (after [12])*

4 Atomic layer deposition (ALD)

In ALD, thin films are built up in cycles in which the surface is exposed to various gas-phase species in alternating, separated doses. In each cycle, a sub-monolayer of a material is deposited. As illustrated in Fig. 13, a typical *ALD cycle* for a simple binary compound consists of four steps: (1) dosing of precursor vapour (precursors see Fig. 5); (2) a purge and/or pump step; (3) dosing of co-reactant, typically involving a small molecule; and (4) a purge and/or pump step. The precursor, and in many cases also the co-reactant, bring elements to the surface that lead to film growth. For the precursor, the element to be deposited is in many cases the metal centre, while for the reactant, it is typically a non-metal such as O, N, S, etc. For ALD, it is vital that the precursor and co-reactants react with the surface in a *self-limiting* way. This means that the precursor molecules and co-reactant species react with surface sites and/or surface chemical groups as long as these are present or accessible; hence the surface reactions eventually saturate and stop. The precursor molecules and co-reactants react neither with themselves nor with the surface groups that they create. In the purge and/or pump steps, the gaseous reaction products that may be generated during the surface reactions, as well as any excess precursor or co-reactant molecules, are removed from the ALD reactor. This is necessary to avoid reactions between precursor and co-reactant molecules directly in the gas phase or on the surface, as this could lead to an undesired CVD component.

4.1 Self-limiting surface reactions

Fig. 13 shows a schematic representation of the self-limiting surface reactions during the two ALD half-cycles. The lower panels show the resulting coverage, or growth per cycle, as a function of exposure or time for that particular step. For sufficient exposure, saturated growth is obtained, while insufficient exposure results in incomplete saturation. For insufficient purging, a CVD component from mixing of the precursor and co-reactant is obtained. The saturation of both ALD half-cycles leads to a characteristic amount of growth per cycle, i.e. GPC measured in nm.

The chemical equations for the surface reactions during the half-cycles might illustrate the ligand exchange reactions in a general form.[9] Considering the growth of a binary material AB from precursor vapour (AX_2) and co-reactant vapour (BY_2), both with two ligands (X and Y, respectively), the reaction during the first half-cycle reads:

$$\text{s-Y}_{(\text{ads})} + \text{AX}_{2(\text{vap})} \rightarrow \text{s-AX}_{(\text{ads})} + \text{XY}_{(\text{g})} \tag{6}$$

's-' indicates the surface with surface groups Y, and XY is the gaseous reaction product. The reaction during the second half-cycle is of the form:

$$\text{s-AX}_{(\text{ads})} + \text{BY}_{2(\text{vap})} \rightarrow \text{s-ABY}_{(\text{ads})} + \text{XY}_{(\text{g})} \tag{7}$$

Note that in other ALD processes the number of ligands and surface groups can differ, resulting in more diverse reaction products. To reach a certain film thickness the two half-cycles (6) and (7) are repeated in an ABAB fashion.

4.2 ALD (temperature) window

For each ALD process certain chemical and physical conditions have to be established in order to obtain self-limiting growth. ALD behaviour which results in a characteristic GPC for a certain process can only be obtained in a specific temperature window. In this regime, the GPC shows a weak or no temperature dependence as shown in Fig. 14. Outside the temperature window, several processes can disrupt the ALD behaviour. At low temperatures, (1) *condensation* of some precursors and co-reactants on the surface can prevent effective purging and lead to an increase in GPC. Alternatively, (2) a *low reactivity* of the molecules with the surface sites due to limited thermal energy can prevent saturation of the reaction and lead to a decrease in growth. At high temperatures, (3) *decomposition* of the precursors or co-reactants can introduce an additional CVD component and lead to an increase in GPC. But also (4) *desorption or etching* effects may occur which remove already deposited material and lead to a decrease in growth.

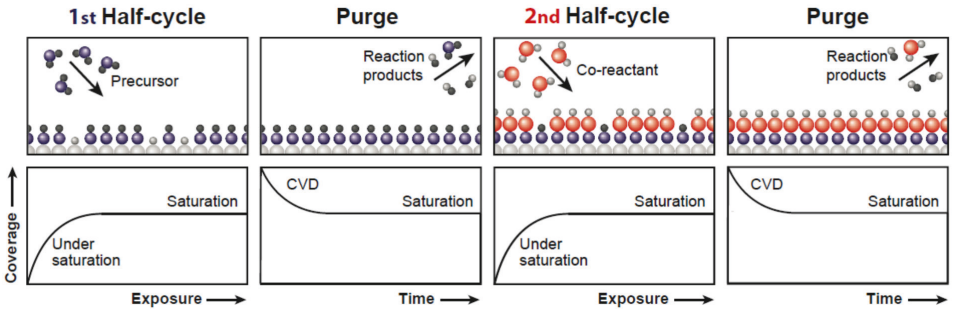


Fig. 13: *A schematic representation of an ALD cycle consisting of two half reactions. The exposures in the first half-cycle (precursor) and second half-cycle (co-reactant) are self-limiting such that the process stops when all available surface sites are occupied. The two half-cycles are separated by purge steps. The half-cycles are repeated in an ABAB fashion to build the film up to the target thickness. (after [9])*

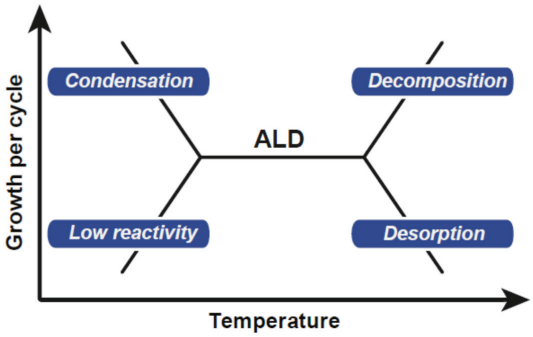


Fig. 14: *Ideal ALD behaviour is characterized by self-limiting growth which shows a weak or no dependence on temperature, while outside the window the ALD behaviour is lost due to one of four disruptive effects. [9]*

4.3 Uniformity and conformality

The surface reaction controlled self-limiting nature of an ideal ALD process affords a constant thickness of the grown film, even in case of local variations of the flux of source species, either at different areas on a substrate or in a three-dimensional structure. The only requirement that needs to be fulfilled is that a sufficient flux reaches all areas. However, for certain cases like for example large wafers with increased surface area due to severe 3-dimensional (3D) pinhole structures, wafer batch processing, and also viscous precursor vapour the requirement for sufficient flux at all areas is not trivial reached. While reactor engineering can improve the uniformity for flux-controlled growth, it cannot typically improve the conformality. The two cases of coverage for surface-controlled and flux-controlled growth are compared in Fig. 15 with respect to the coverage of relatively large planar areas (uniformity) and the coverage of a 3D structure (conformality).

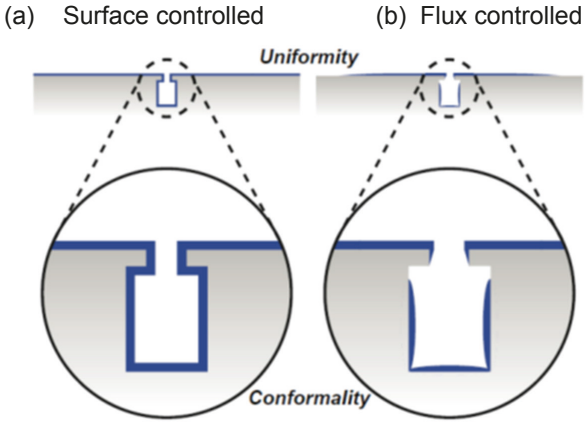


Fig. 15: *The uniformity and conformality for (a) surface-controlled and (b) flux-controlled deposition growth. (taken from [9])*

4.4 Advanced ALD process schemes

Besides the simple ALD process with two alternating half-cycles (Fig. 13), there are many additional methods of exploiting ALD-type processes by using more steps as shown in Fig. 16. In a multistep process (ABC cycle) additional steps can be used to change the process to widen the temperature window or achieve different material properties. *Supercycles* a used to grow alloy, doped or multilayer films of specific elemental mixtures.

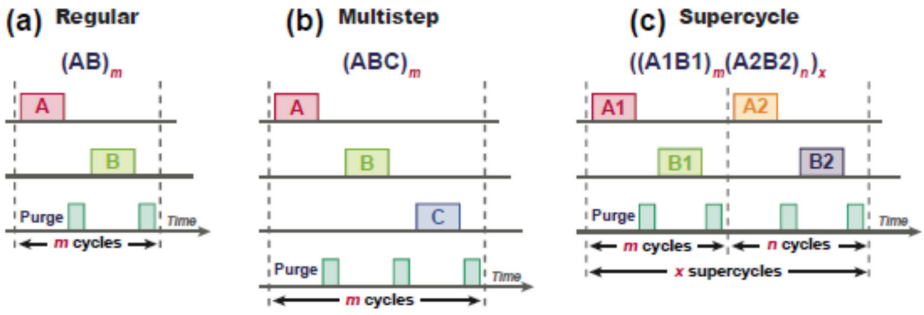


Fig. 16: *A schematic representation of the various steps in (a) a regular ALD process, (b) a multistep process and (c) a supercycle. (after[9])*

4.5 ALD co-reactants

For ALD processes, in general, the same precursors can be used as for MOCVD processes provided that the additional requirements discussed in section 2.1 (see Fig. 4) are fulfilled. However, in ALD a suitable co-reactant is needed for completing the second half-cycle of the ALD process (see Fig. 13). The main purpose of the co-reactant molecule is to react cleanly with surface ligands, thereby adding a second component to the film (where required), and to reform the original surface groups. Table 1 shows a list arranged by Knoops et al. [9] of the most common ALD materials alongside common nonmetal sources (or co-reactants) for those materials. Additional information and references can be obtained from the comprehensive review by Miikulainen et al [8]. In general, co-reactants are volatile, small molecules, such as elements (H_2 or O_2), hydrides (H_2O, NH_3 etc.), or alkyl compounds (BEt_3 or $AsMe_3$). For metal oxide and nitride deposition the most popular co-reactants are H_2O and NH_3, respectively. Analogous for the deposition of metal sulfides, selenides, and tellurides, either hydrides or alkyl compounds are typical. The deposition of pure metals has also been made possible by a wide range of co-reactants, including H_2 and even O_2. In general for metals, the co-reactant should be a reducing agent such that the pure ligand-free metal remains. The use of O_2 is restricted to the ALD of catalytic metals that are more thermodynamically stable than their respective oxide. With decreased growth temperatures, the reactivity and the difficulty to purge from the reactor become serious issues, especially for H_2O which strongly absorbs to surfaces. Higher reactivity at low temperature can be obtained for species with relatively short lifetimes like ozone, O_3, and plasma species as examples.

Table 1: *Typical ALD materials and the common (non-metallic) co-reactants used, where R replaces H or any alkyl or aryl group [9].*

Material	Common Co-reactants
Metal oxides	H_2O, H_2O_2, ROH, R(CO)H, R(CO)OH, R_2CO, O_2, O_3, O_2 plasma, O radicals, $M(OR)_x$, CO_2, N_xO_y, air
Metal nitrides	NH_3, NR_3, R_2NNR_2, NH_3 plasma, H_2 plasma, N_2 plasma, H_2-N_2 plasma, NH_x radicals
Metal carbides	C_xH_y, BR_3
Metal phosphides	PR_3, $POCl_3$, $P(OR)_3$, $PO(OR)_3$, $P(NR_2)_3$
Metal arsenides	AsH_3, AsR_3, $As(NR_2)_3$
Metal sulfides	H_2S, S_2R_2
Metal selenides	H_2Se, R_2Se, $Se(SiR_3)_2$
Metal tellurides	H_2Te, $Te(SiR_3)_2$
Metal fluorides	HF, MF_x
Pure element (metal)	H_2, H_2 plasma, NH_3, NH_3 plasma, H_2 plasma, N_2 plasma, H_2-N_2 plasma, NH_x radicals, O_2, O_2 plasma, O radicals, Si_xH_y, formalin

Because of the short lifetimes these co-reactants must be produced in situ, rather than being obtained from a chemical supplier. Their production requires an application of energy e.g. in the form of an electrical discharge, thermal cracking, or photodissociation. Processes involving such extra energy are referred to as energy-enhanced ALD. The benefit of these is that a high reactivity is obtained, and long purge times are often not necessary. After this general introduction into ALD concepts, in the next chapter we will cover chemistries of specific ALD processes with focus on specific materials useful for memristor applications.

5 ALD chemical reactions

A variety of reaction chemistries are applicable to ALD processes [9, 13]:

- *Ligand-exchange*: a reaction between surface groups and ligands on precursors / co-reactants, where groups are exchanged leading to volatile reaction products. These are sequential *condensation* and *hydrolysis* reactions.
- *Dissociation*: a reaction in which the precursor / co-reactant dissociate into several adsorbed species on the surface without releasing reaction products into the vapour phase.
- *Association*: the bonding of an intact precursor / co-reactant with the surface without release of ligands (e.g. hydrogen-bonding).
- *Combustion/ nitridation*: an oxidizing co-reactant can combust surface groups and replace them with an oxidized surface, analogously the co-reactant can nitridate.
- *Abstraction*: a co-reactant can remove ligands and release reaction products, without leaving behind fragments from the co-reactant, for instance, in conjunction with the reduction of the metal centre.
- *Reduction*: the metal centre can be reduced to a lower oxidation state by the co-reactant.

In the following, a few selected examples will be given. Further reactions and references are given in review articles like for example [8].

5.1 Al_2O_3 ALD

The best understood ALD chemistry is that of Al_2O_3. Especially the 'TMA/water' process where trimethyl aluminium (TMA, correctly a dimer below 70°C, i.e. $[Al(CH_3)_3]_2$) is used as the Al-precursor and water vapour as the co-reactant is a prominent example for the ligand exchange mechanism [14]. The 'TMA/O_2 plasma' process demonstrates how an O_2 plasma undergoes combustion-like mechanisms with an organically terminated surface [9]. Other Al_2O_3 ALD processes involve halides (e.g. Al_2Cl_6) and alkoxides (e.g. $[Al(OiPr)_3]_4$).[8] Recently, a heteroleptic Al-precursor, dimethyl aluminum isopropoxide (DMAI), has been introduced as a non-pyrophoric alternative to TMA [15, 16].

TMA/water

The two ALD half-cycles of the TMA/water process are good examples for Brønsted acid-base ligand-exchange reactions, i.e. alternating *condensation* and *hydrolysis* reactions between the surface groups and the incoming precursor(s) releasing volatile reaction products. During the TMA half-cycle (condensation, n = 1, 2) *hydrogen atoms transfer* from surface hydroxyl groups to methyl ligands of the TMA.

$$n\ \text{s-O}\mathbf{H}_{(\text{ads})} + \tfrac{1}{2}\left[\text{Al}(\text{CH}_3)_3\right]_{2\,(\text{vap})} \rightarrow \text{s-O}_n\text{Al}(\text{CH}_3)_{3-n\,(\text{ads})} + n\,\mathbf{H}\text{CH}_{3(\text{g})} \qquad (8)$$

During the water half-cycle (hydrolysis) *hydrogen atoms transfer* from water to surface-bound methyl ligands.

$$\text{s-Al}(\text{CH}_3)_{(\text{ads})} + \mathbf{H}\text{OH}_{(\text{vap})} \rightarrow \text{s-Al}\text{OH}_{(\text{ads})} + \mathbf{H}\text{CH}_{3\,(\text{g})} \qquad (9)$$

Thus, the surface hydroxyl group (s-OH) acts as a Brønsted acid by donating a proton to a methyl (CH_3) ligand on the TMA, which is the Brønsted base. This transfer results in the formation of a surface O–Al bond and the release of methane (CH_4) as a reaction product. Additionally, TMA may bind either via one or two methyl groups (mono- or bifunctional binding). Fig. 17 shows the corresponding graph of growth per cycle (GPC) as a function of temperature, representing well a *thermal* ALD process (see Fig. 14). At temperatures below 150 °C, the GPC drops with decreasing temperature due to a lower reactivity. In the temperature window even the TMA/water process doesn't show ideal ALD behaviour, i.e. a constant GPC. Instead, a slight drop is observed with increasing temperature which stems from a reduction in the density of reactive s-OH groups [14].

TMA/O_2 plasma

For the TMA/O_2 plasma ALD process, the first half-cycle can essentially be considered the same as in the thermal process (see eq. (8)), whereby incoming TMA molecules react with s-OH groups via hydrogen atoms transfer. The O_2 plasma half-cycle (*combustion*) involves a *transfer of oxygen atoms*. O_2 plasma species (analogue to ozone [17]) oxidize the surface-bound methyl groups to carbon dioxide and water, and also transfer oxygen atoms to the surface bound aluminium, recreating a hydroxylated surface ready to begin the next cycle.

$$\text{s-Al}(\text{CH}_3)_{(\text{ads})} + 4\mathbf{O}_{(\text{plasma})} \rightarrow \text{s-Al}\mathbf{O}\text{H}_{(\text{ads})} + \text{C}\mathbf{O}_{2\,(\text{g})} + \text{H}_2\mathbf{O}_{(\text{vap})} \qquad (10)$$

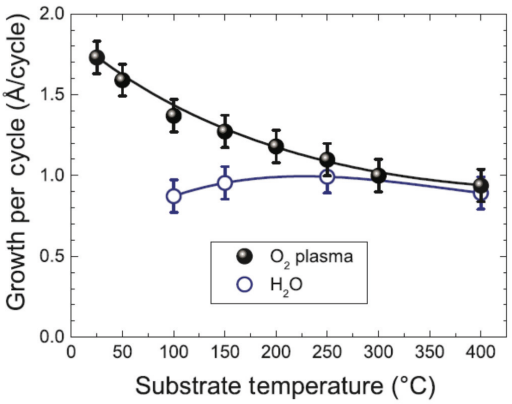


Fig. 17: *Variation in growth per cycle (GPC) as a function of substrate temperature for the ALD of Al_2O_3 from TMA, $[Al(CH_3)_3]_2$ and water vapour or O_2 plasma, repectively. [9]*

As plasmas are highly reactive, Al_2O_3 films synthesized by plasma-enhanced ALD tend to have a higher GPC (Fig. 14), higher density and lower carbon, hydrogen, and (excess) oxygen contents than films deposited using water or (even) ozone as the co-reactant. This is particularly the case at low substrate temperatures [18].

5.2 HfO_2 ALD

ALD ultrathin hafnium oxide films, which are now being evaluated for ReRAM applications, have become famous as dielectric layers in high-k metal gate field effect transistors [6]. In industry, halide-based precursors, in particular $HfCl_4$, are favoured because of the reactivity (with surface OH) and stability of the compound. Other common precursors are alkylamide-based ones, like for example the heteroleptic precursor tetra-ethyl-methyl-amino-Hf (TEMAH, $Hf(NEtMe)_4$) or the homoleptic one TDEAH ($Hf(NEt_2)_4$) which are widely reported in industrial and academic research and development [19, 20].

$HfCl_4$/water

A prominent example for halide-based ALD processes is the reaction of $HfCl_4$ with water. The half-cycles of condensation (eq. (11)) and hydrolysis (eq. (12)) appear to be similar to the TMA/water process [14].

$$n\ \text{s-O}\mathbf{H}_{(\text{ads})} + \text{HfCl}_{4\,(\text{vap})} \rightarrow \text{s-O}_n\text{HfCl}_{4-n\,(\text{ads})} + n\,\mathbf{H}\text{Cl}_{(\text{g})} \tag{11}$$

$$\text{s-HfCl}_{(\text{ads})} + \mathbf{H}_2\text{O}_{(\text{vap})} \rightarrow \text{s-HfO}\mathbf{H}_{(\text{ads})} + \mathbf{H}\text{Cl}_{(\text{g})} \tag{12}$$

In addition, $HfCl_4$ has an affinity to oxygen bridges, i.e. Hf–O–Hf. The $HfCl_4$ can add across an oxygen bridge (eq. (13)), and remaining chlorides can undergo hydrolysis with water (eq. (14)). In essence, the oxygen bridges can serve as reactive surface sites.[21]

$$\text{s-Hf(-O-)}_2\text{Hf}_{(\text{ads})} + \text{HfCl}_{4\,(\text{vap})} \rightarrow \text{s-Hf(Cl)-O-Hf(OHfCl}_3)_{(\text{ads})} \tag{13}$$

$$\text{s-Hf(Cl)-O-Hf(OHfCl}_3)_{(\text{ads})} + \text{H}_2\text{O}_{(\text{vap})} \rightarrow \text{s-Hf(-O-)}_2\text{Hf(-O-)}_2\text{Hf}_{(\text{ads})} + 4\text{HCl}_{(\text{g})} \tag{14}$$

$HfCl_4$/ozone

The deposition of HfO_2 from $HfCl_4$ can also be performed with ozone (O_3) as the co-reactant.[22] Assuming only oxygen bridges or surface oxide were present, the $HfCl_4$ could either react by adding across an Hf–O–Hf bridge (eq. (13)) or simply by binding to surface oxygen species (eq. (11)). For the reaction between incoming ozone and the surface chloride species it has been proposed that the ozone oxidizes the surface chloride species to chlorine gas (eq. (15)) while forming a Hf–O bond via an *oxygen transfer* reaction.

$$\text{s-HfCl}_{2(\text{ads})} + 2\mathbf{O}_{3(\text{g})} \rightarrow \text{s-Hf}\mathbf{O}_{2(\text{ads})} + 2\mathbf{O}_{2(\text{g})} + \text{Cl}_{2(\text{g})} \quad (15)$$

TDMAH/water

Another example for amide-based ALD processes is the reaction of tetrakis (dimethylamido) hafnium, $Hf(NMe_2)_4$, with water to make HfO_2.[19] The overall reaction is:

$$\text{Hf(NMe}_2)_4 + 2\,\text{H}_2\text{O} \rightarrow \text{HfO}_2 + 4\,\text{HNMe}_2 \quad (16)$$

Chemisorption of the Hf precursor occurs by hydrogen transfer to the dimethylamido ligands to release dimethylamine gas (eq. (17)). In the second half-cycle (eq. (18)) hydrogen is transferred from water to the remaining surface-bound diethylamide ligands.

$$2\ \text{s-O}\mathbf{H}_{(\text{ads})} + \text{Hf(NMe}_2)_{4(\text{vap})} \rightarrow \text{s-O}_2\text{Hf(NMe}_2)_{2(\text{ads})} + 2\,\mathbf{H}\text{NMe}_{2(\text{g})} \quad (17)$$

$$\text{s-O}_2\text{Hf(NMe}_2)_{2(\text{ads})} + 2\,\mathbf{H}\text{OH}_{(\text{vap})} \rightarrow \text{s-O}_2\text{Hf(OH)}_{2(\text{ads})} + 2\,\mathbf{H}\text{NMe}_{2(\text{g})} \quad (18)$$

The aforementioned chemistries for the Al_2O_3 and HfO_2 processes demonstrate how ALD manifests itself in real reaction mechanisms. However, there are many alternative ALD chemistries, which can seem increasingly complex, like for example the ALD processes for phase change materials i.e. higher chalcogenides, shown in the following.

5.3 GeTe ALD

As an example, the ALD process for germanium telluride (GeTe) which is one member of the GST series of phase change materials [23] should be discussed. *Chlorine atom transfer* plays a key role in a process for ALD of GeTe.[24] Chlorine atoms on a previously chlorinated germanium surface transfer to trialkylsilicon groups released from the bis(trialkylsilyl)tellurium precursor vapour (eq. (19)). During the second half-cycle (eq. (20)) chlorine atoms on the Ge precursor vapour remove trialkylsilyl groups. One example for GeTe growth should be mentioned here. The main driving force for theses reactions is the formation of the stronger Si-Cl bonds after breaking the weaker Ge-Cl bond. A wide variety of selenides and tellurides have been made by similar reactions. [13]

$$2\ \text{s-Ge}\mathbf{Cl}_{(\text{ads})} + (\text{R}_3\text{Si})_2\text{Te}_{(\text{vap})} \rightarrow \text{s-Ge-Te-SiR}_{3(\text{ads})} + \text{R}_3\text{Si}\mathbf{Cl}_{(\text{g})} \quad (19)$$

$$\text{s-Ge-Te-SiR}_{3(\text{ads})} + \text{Ge}\mathbf{Cl}_{2(\text{vap})} \rightarrow \text{s-Ge-Te-GeCl}_{(\text{ads})} + \text{R}_3\text{Si}\mathbf{Cl}_{(\text{g})} \quad (20)$$

6 Growth control

6.1 Uniformity and conformality

Real ALD processes rely on surface chemistry, and therefore chemical side effects can be present which hinder the perfect uniformity and conformality of the grown film over the entire surface of the sample. Among these are: (1) *decomposition* of the precursor in consequence of a too long residence time in the reactor or, especially, in severe pinhole structures on the substrate; (2) *etching*, which for instance, can be an issue with metal halide precursors; (3) *surface poisoning*, a chemical side effect in which the reaction products influence the process by competing with the precursor for the surface groups created during the co-reactant exposure (see Fig. 18); (4) *process interaction*, which has to be considered when using different processes in sequence, for instance in a supercycle (see for example the growth of multi-component oxides in [25]).

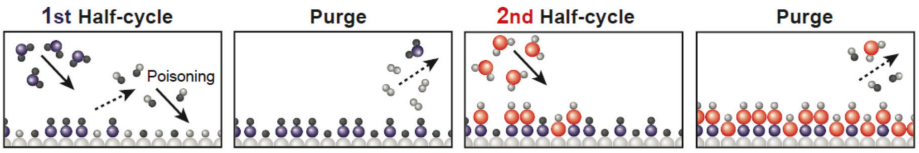


Fig. 18: *A schematic representation of surface poisoning during ALD. The reaction-products during the first half-cycle can also react with the initial surface and, therefore, compete with the precursor for surface sites, which can lead to nonuniform deposition. [9]*

However, by means of precursor chemistry and process modification, excellent conformality has been achieved for many ALD processes as can be seen, for instance, in Fig. 19 for the growth of a $Ge_2Sb_2Te_5$ film into trenches [24].

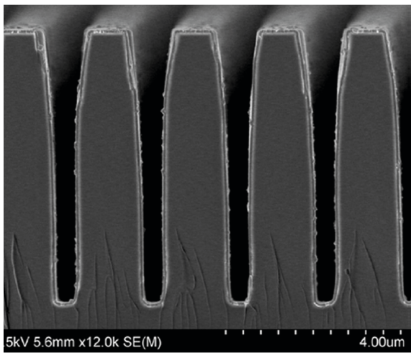


Fig. 19: *SEM cross-sectional image of conformal $Ge_2Sb_2Te_5$ ALD film in trenches. [24]*

6.2 Early stages of growth – towards ultrathin films

The substrate material and its surface groups can lead to a growth delay or growth enhancement at the early stages of ALD growth which are especially important for growth control of ultrathin films of a few nanometer thickness. Fig. 20 shows, schematically, the effect on the average growth per cycle (GPC) for different scenarios. An immediate constant GPC is expected for an ideal process, while an accelerated or delayed growth will lead to higher or lower thickness, respectively. Issues with nucleation of the material to be deposited on a surface can hinder growth of closed thin films, but can also be exploited to obtain controlled growth of islands or even nanoparticles on surfaces [9]. The thermal ALD process of HfO_2 using $HfCl_4$ and H_2O is an example of growth delay due to a limited number of OH groups on the initial surface.[14] The growth proceeds through the formation of islands on isolated reactive sites, which slowly coalesce during consecutive cycles, after which normal growth is obtained. When the sum of the surface energy of the material to be deposited and of the interface energy between film and substrate is considerably higher than the surface energy of the substrate, de-wetting of the film can occur, leading to island or nanoparticle formation. In principle, it is possible to deposit films of thickness down to atomic layers in the case of a 'Frank – van der Merve'-type growth, i.e. for substrates which are "wetted" by the deposited material. This illustrates that, in some cases, ALD can be used to deposit either nanoparticles or closed films using the number of cycles for control.

Besides the common goal of optimizing conditions and processes in order to reduce the dependency on the starting surface properties, the opposite is also desired in some cases. Here the process can be developed such that *selective growth* occurs [26]. In this case, there is controlled growth on certain materials, while there is no growth on other materials. Such a process can be exploited in order to achieve self-assembly or, for instance, a mask-less process [5]. Supercycles can be used to control the dopant introduction in a material. Al-doped TiO_2 has been deposited by choosing the metal precursor of the Al_2O_3 process, the lateral dopant distance can be changed while the dopant distance perpendicular to the substrate is controlled by the number of TiO_2 cycles [27].

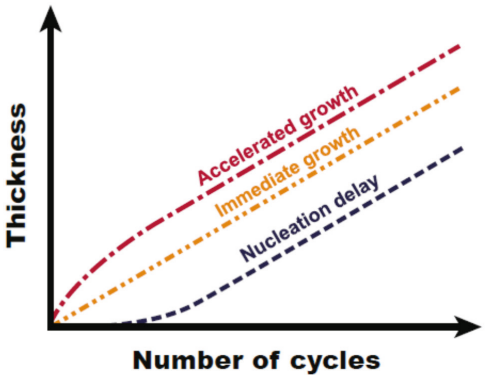


Fig. 20: *Film thickness as a function of number of cycles during an ALD process. Three kinds of nucleation behaviour of the ALD films can typically be distinguished during the first few cycles. ALD process with: accelerated growth, immediate constant GPC (ideal process), and (pronounced) nucleation delay. [9]*

6.3 Growth at low temperatures

Compared to CVD, ALD can typically achieve excellent material properties at lower deposition temperatures around 300 °C. However, at even lower temperatures (25–100 °C), thermal ALD processes can have low material purity, low GPC values, or long cycle times [28]. Energy-enhanced ALD (e.g., plasma ALD), can generally be used to overcome this deficiencies because of its higher reactivity. Using plasma ALD, several oxides have been deposited at temperatures down to room temperature such as Al_2O_3, TiO_2, and SiO_2 as shown in Fig. 21 [29]. All plasma enhanced ALD processes and the TMA/O_3 ALD process showed a linear increase in thickness with the number of ALD cycles.

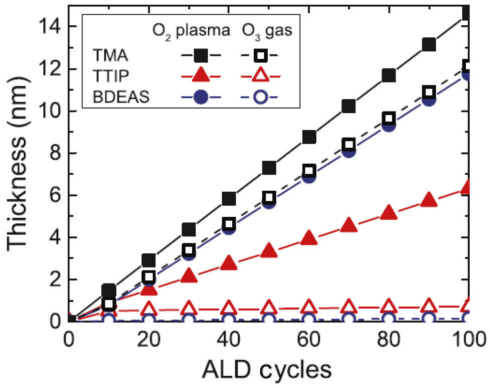


Fig. 21: *The increase in film thickness as a function of ALD cycles for Al_2O_3, TiO_2, and SiO_2 performed at room temperature using the precursors TMA, $Ti(OiPr)_4$ (TTIP), and $SiH_2(NEt_2)_2$ (BDEAS), respectively. [29]*

6.4 Microstructure and phase control

Whereas CVD techniques allow even for the growth of epitaxial layers due to the higher thermal budget involved, the microstructures of ALD films can still be adjusted in the range from amorphous to polycrystalline. Regarding application in resistive switching cells the control of the microstructure and phase of the films is of superior importance because this structure determines the electronic and ionic conductivity of the ReRAM cell.

For HfO_2 based cells different strategies are followed: (1) completely amorphous films guarantee a high reproducibility of the devices while (2) polycrystalline films exhibit an inherent inhomogeneity which might lead to larger device-to-device variations, especially for decreasing device size. In contrast, grain boundaries in polycrystalline HfO_2 films have been utilized as weak spots which enable electroforming of the ReRAM cells at reduced voltages compared to amorphous layers [30]. For TiO_2 based cells it has been analysed that the conductive filament in the ReRAM cell consists of a Magnéli-type phase [31]. Utilizing a liquid injection ALD technique it has been demonstrated that even the growth of titanium sesquioxide phases is possible by means of process control [32]. In general, for TiO_2, higher temperature promotes crystalline growth. However, it is not totally clear whether this is a direct consequence or an indirect effect associated with the decreasing

impurity content with increasing temperature. Besides the temperature other factors are important, which comprise effects of reactants, impurities, plasma enhancement, the substrate, and the film thickness. The comprehensive summary of the crystallinity of metal oxide films grown by two-reactant ALD processes and of chalcogenide films is provided by the review of Miikulainen et al. [8] to which we refer here for further details. The authors identified several trends regarding the crystallinity of inorganic ALD films, whereas not all ALD processes follow these trends. The higher the ALD temperature, the thicker the film, and the purer the resulting material, the more likely the resulting film will be crystalline. The use of plasma enhancement increases the probability of depositing a crystalline film. Polycrystalline ALD films typically consist of columnar grains whereas the extent of randomness in the orientation depends on the substrate and temperature. The crystalline grain size is typically related to the film thickness, although in some cases it can be order(s) of magnitude larger than the film thickness.

7 Summary

Chemical vapour deposition techniques have been introduced with a focus on processes used for the growth of the actively switching layers in memristive devices, either redox based ReRAM devices or phase change systems. As a consequence of the aggressive trend in miniaturization, dimensions of the actively switching areas are continuously decreasing to about 1000 nm^3. This trend sets strong demands on the film deposition method which has to provide atomic precision of film thickness and composition together with a high conformality and at low growth temperatures. Atomic layer deposition is the only technique which can fulfill the above requirements, and ALD has been therefore introduced to some broader extent. The low deposition rate compared to CVD is often named as a disadvantage of ALD but with continued scaling this will become in fact an advantage because even thinner films are typically required for smaller technology nodes and low rates but excellent control over thickness are beneficial. Therefore, ALD might play an important role in future high density ReRAM and PCRAM fabrication. Despite all the successful ALD precursors and processes have been discovered so far, there still remain many areas for improvement in the field of compositions, purity, phase composition, structure, morphology, and control of the defect states of materials produced by ALD.

References

[1] Waser, R., Bruchhaus, R. & Menzel, S. (2012) Redox-based Resistive Switching Memories, Chapter 30, 683-710, in Waser, R. Nanoelectronics and Information Technology (3rd edition), Wiley VCH.

[2] Govoreanu, B., Kar, G.S., Chen, Y-Y., Paraschiv, V., Kubicek, S., Fantini, A., Radu, I.P., Goux, L., Clima, S., Degraeve, R., Jossart, N., Richard, O., Vandeweyer, T., Seo, K., Hendrickx, P., Pourtois, G., Bender, H., Altimime, L., Wouters, D.J., Kittl, J.A. & Jurczak, M. (2011) 10x10nm2 Hf/HfOx Crossbar Resistive RAM with Excellent Performance, Reliability and Low-Energy Operation, IEEE International Electron Devices Meeting - IEDM '11, 31.6.1-4.

[3] Hsu, C.-W., Wang, I-T., Lo, C.-L., Chiang, M.-C., Jang, W.-Y., Lin, C.-H. & Hou, T.-H. (2013) Self-Rectifying Bipolar TaOx/TiO2 RRAM with Superior Endurance over 1012 Cycles for 3D High-Density Storage-Class Memory, 2013 Symposium on VLSI Technology Digest of Technical Papers, T166-167.

[4] Jones, A. C. & Hitchman, M. L. (2008) Chemical Vapour Deposition: Precursors, Processes and Applications, Royal Society of Chemistry, Thomas Graham House, Cambridge, England.

[5] Pinna, N. & Knez, M. (2011) Atomic Layer Deposition of Nanostructured Materials, WILEY-VCH.

[6] Hwang, C. S. (2014) Atomic Layer Deposition for Semiconductors, Springer, New York, USA.

[7] Devi, A. (2013) 'Old Chemistries' for new applications: Perspectives for development of precursors for MOCVD and ALD applications, Coord. Chem. Rev. 257 3332-3384.

[8] Miikkulainen, V., Leskelä, M., Ritala, M. & Puurunen, R. L. (2013) Crystallinity of inorganic films grown by atomic layer deposition: Overview and general trends J. Appl. Phys. 113 021301/1-100.

[9] Knoops, H.C.M., Potts, S.E., Bol, A.A. & Kessels, W.M.M. (2015) Atomic layer deposition, Chapter 27, 1101-1133, in Handbook of Crystal Growth: Thin Films and Epitaxy, Elsevier.

[10] Ehrhart, P. (2003) Film Deposition Methods, Chapter 8, 199-221, in Waser, R. Nanoelectronics and Information Technology (2nd edition), Wiley VCH.

[11] Hwang, C.S., Park, J., Hwang, D.S. & Yoo, C.Y. (2001) J. Electrochem. Soc. 148, G636.

[12] Schumacher, M., Baumann, P. K., Lindner, J., Lohe, C., Weber, U., Ramanathan, S., Karim, Z., Londergan, A. R. & Seidel, T.E. (2005), Atomic vapor deposition (AVD) for next generations of advanced semiconductor devices, 208th ECS Meeting, Abstract 487.

[13] Gordon, R. G. (2014) ALD Precursors and Reaction Mechanisms, Chapter 2 in Hwang, C. S. Atomic Layer Deposition for Semiconductors, Springer, New York.

[14] Puurunen, R. L. (2005) Surface chemistry of atomic layer deposition: a case study for the trimethylaluminum/water process, J. Appl. Phys. 97, 121-301.

[15] Potts, S.E., Dingemans, G., Lachaud, C. & Kessels, W. M. M. (2012) J. Vac. Sci. Technol. A 30, 21505/1.

[16] Zhang, H., Aslam, N., Reiners, M., Waser, R. & Hoffmann-Eifert, S. (2014) Atomic layer deposition of TiOx/Al2O3 bilayer structures for resistive switching memory applications, Chemical Vapor Deposition 20, 282-290.

[17] Heil, S.B.S., van Hemmen, J.L., van de Sanden, M.C.M. & Kessels, W.M.M. (2008) J. Appl. Phys. 103,103302.

[18] Potts, S.E., Keuning, W., Langereis, E., Dingemans, G., van de Sanden, M.C.M. & Kessels, W,M,M. (2010) Low temperature plasma-enhanced atomic layer deposition of metal oxide thin films, J. Electrochem. Soc. 157,66–74.

[19] Hausmann, D. M., Kim, E., Becker, J. & Gordon, R. G. (2002) Atomic layer deposition of hafnium and zirconium oxides using metal amide precursors, Chem. Mater. 14, 4350.

[20] Niinistö, J., Kukli, K., Kariniemi, M., Ritala, M., Leskelä, M., Blasco, N., Pinchart, A., Lachaud, C., Laaroussi, N., Wang, Z. & Dussarrat, C. (2008) J. Mater. Chem. 18, 5243.

[21] Aarik, J., Aidla, A., Kiisler, A.A., Uustare, T. & Sammelselg, V. (1999) Thin Solid Films 340,110–6.

[22] Delabie, A., Swerts, J., van Elshocht, S., Jung, S.H., Raisanen, P.I., Givens, M.E., et al. (2011) J. Electrochem. Soc. 158, D259–63.

[23] Lencer, D., Salinga, M., Grabowski, B., Hickel, T., Neugebauer, J. & Wuttig, M. (2008) A map for phase-change materials, Nat. Mater. 7, 972-977.

[24] Pore, V., Hantanpää, T., Ritala, M. & Leskelä, M. (2009) Atomic layer deposition of metal tellurides and selenides using alkylsilyl compound of tellurium and selenium, J. Am. Chem. Soc. 131, 3478.

[25] Hoffmann-Eifert, S. & Watanabe, T. (2014) FeRAM, Chapter 6 in Hwang, C. S. Atomic Layer Deposition for Semiconductors, Springer, New York.

[26] Lu, J., Elam, J.W. & Stair, P.C. (2013) Acc. Chem. Res. 46, 1806–15.

[27] Kim, S.K., Choi, G-J., Lee, S.Y., Seo, M., Lee, S.W., Han, J.H., Ahn, H-S., Han, S. & Hwang, C. S. (2008) Al-doped TiO2 films with ultralow leakage currents for next generation DRAM capacitors, Adv. Mater. 20, 1429.

[28] Groner, M.D., Fabreguette, F.H., Elam, J.W. & George S.M. (2004) Chem. Mater. 16, 639–45.

[29] Potts, S.E., Profijt, H.B., Roelofs, R. & Kessels, W.M.M. (2013) Chem. Vap. Deposition 19,125–33.

[30] Lanza, M., Bersuker, G., Porti, M., Miranda, E., Nafria, M. & Aymerich, X. (2012) Resistive switching in hafnium dioxide layers: Local phenomenon at grain boundaries, Appl. Phys. Lett. 101, 193502.

[31] Kwon, D.-H., Kim, K. M., Jang, J. H., Jeon, J. M., Lee, M. H., Kim, G. H., Li, X.-S., Park, G.-S., Lee, B., Han, S., Kim, M. & C. S. Hwang (2010) Atomic structure of conducting nanofilaments in TiO2 resistive switching memory, Nat. Nanotechnol. 5, 148-153.

[32] Reiners, M., Xu, K., Aslam, N., Devi, A., Waser, R. & Hoffmann-Eifert, S. (2013) Growth and crystallization of TiO2 thin films by atomic layer deposition using a novel amido guanidinate titanium source and tetrakis-dimethylamido-titanium, Chem. Mater. 25, 2934-2943.

B2 Physical Deposition Techniques

Regina Dittmann
Forschungszentrum Jülich (PGI-7), Germany

Contents

1 Introduction **2**

2 Fundamentals of Thin Film Deposition **3**
2.1 Gas Kinetics and Thermodynamics 3
2.2 Relevant processes on the substrate surface 4
2.3 Growth Modes 5
2.4 Strain relaxation 7

3 Substrates and its surface termination **8**

4 In-situ characterization / RHEED **9**

5 Molecular beam epitaxy **11**
5.1 Evaporation techniques 11
5.2 MBE systems 12
5.3 Adjusting the stoichiometry in oxide thin films 13
5.4 Molecular beam epitaxy of GST 14

6 Sputtering **15**
6.1 Basic principle 15
6.2 Sputtering techniques 16
6.3 Modification of thin film morphology 16
6.4 Stoichiometry adjustment during sputter growth 17

7 Pulsed laser deposition **19**
7.1 Basic principle 19
7.2 Stoichiometry adjustment during PLD growth 20

8 Summary **23**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

In order to employ materials with memristive properties, e.g. resistively switching oxides or higher chalcogenides in future non-volatile memories or neuromorphic circuits, it is inevitable to fabricate them in thin film form. Complementary to the chemical vapour deposition techniques addressed in the contribution B1, we will give an overview over the most prominent physical vapour deposition techniques, namely molecular beam epitaxy (MBE), sputtering and pulsed laser deposition (PLD), which have been employed for the growth of oxide and higher chalcogenide thin films. The generic working principle of all physical deposition techniques is that either the whole compound or its the single components are transferred from a solid state source to the vapour phase and subsequently condensed as thin film on an appropriate substrate. Figure 1 summarises the different possibilities to transfer solid material to the vapour phase. The most intuitive way is to heat and thereby vaporise the material. Besides this, it is common for a large variety of materials, in particular for those with high melting points, to induce the melting process by e-beam bombardment.

Instead of going through an equilibrium vapour phase, methods like sputtering or pulsed laser deposition induce the ejection of particles from the surface under non-equillibrium conditions. In the case of sputtering, the ejection of particles is induced by ion bombardment of a compound target. In case of pulsed laser deposition (PLD), laser light induces an excitation and subsequent ejection of particles from the target surface. Although laser irradiation might additionally induce local heating, it is advantageous for most routes of PLD growth to minimize the transformation of laser light into phonons during the target ablation process.

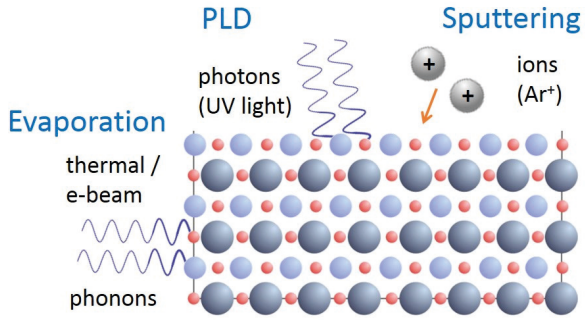


Fig. 1: *Schematic overview over the different methods to transfer solid state to vapor*

All relevant memristive materials have in common that they consist of multiple components. Therefore, the adjustment of the thin film stoichiometry is a key issue for all materials since it has crucial impact on the electronic transport as well as on the memristive behaviour.

In ionic materials like oxides, point defects on the anion and the cation sites act as donors and acceptors, respectively, and have therefore significant influence on the electronic transport. Furthermore, ionic motion as prerequisite for resistive switching in oxides can be considerable enhanced by the presence of point defects. In particular the oxygen stoichiometry is a key parameter to control the switching properties of oxide thin films.

In phase change materials such as Ge-Sb-Te (GST) alloys, the exact composition significantly determines the electrical and optical properties as well as the crystallization speed, which are of key relevance for memory applications.

For oxide thin films, crystallinity is not a prerequisite for stable resistive switching and amorphous thin films are a promising approach for the fabrication of CMOS compatible highly integrated devices. However, single crystalline model systems offer the possibility to study the basics switching mechanism and to clarify the role of certain types of point and extended defects in thin film devices.

Phase change thin films such as GST are typically grown in partially as amorphous layers and are subsequently annealed in order to be crystallized. However, in order to study the details of the structural properties of the different alloy compositions, it is highly advantageous to have access to crystalline thin film model systems. Furthermore, a new concept of phase change memories has been developed recently which requires high quality, crystalline GeTe-Sb_2Te_3 superlattices.

Therefore, the epitaxial growth of oxides as well as phase change materials is highly interesting and according to the complexity of the materials a highly challenging field of research.

In this contribution, we will first present a short overview over the basic mechanisms relevant for physical vapour deposition techniques and thin film growth in general. Afterwards, we will explain the working principle of MBE, sputtering and PLD. For all three techniques, we will give examples how the adjustment of stoichiometry and/or crystallinity is addressed within the different classes of materials.

2 Fundamentals of Thin Film Deposition

2.1 Gas Kinetics and Thermodynamics

For all vapour deposition techniques, the residual gas pressure in the system is one of the basic parameters to be controlled during film deposition. As the residual gas atoms may collide with the depositing species or hit the growing surfaces they may thus be incorporated in the film. In order to guarantee sufficient oxygenation of the thin films during growth, oxides are often deposited under a considerable oxygen flow. In that case, the phase diagrams of the specific compounds have to be taken into account for the selection of the appropriate oxygen gas environment.

For the simplest assumption that the gas atoms may be considered as not interacting masses with a Maxwell velocity distribution, we obtain the mean free path length, λ, of the atoms or molecules.

$$\lambda = \frac{1}{\sqrt{2\pi} N d^2} = \frac{k_B T}{\sqrt{2\pi} p d^2} \tag{1}$$

With d = molecular diameter, N = concentration of the gas. With the law of the ideal gas: $N = p/k_B T$, k_B = Boltzmann constant.

As can be seen from equation 1, the mean free path depends inversely proportional to the pressure p. For the example of air molecules we obtain at room temperature a free path length which is of the order of a typical distance from source to substrate of about 20 cm at a pressure of 0.5 10^{-3} mbar. In order to obtain an undisturbed flux of particles during evaporation, a background pressure below 10^{-6} mbar is required. More critical is the number of residual gas atoms which hit the growing surface and limit the purity of the film if they are incorporated. This number can be expressed as

$$N_i = p_i \sqrt{\frac{1}{2\pi k_B m_i T}} \tag{2}$$

m_i = atomic or molecular mass. Assuming a sticking coefficient of unity, the incorporation of residual gas atoms may be expressed in terms of monolayers and this growth rate may be rather high compared to a typical growth rate of an epitaxial film of one monolayer / s. Hence, for clean films ultrahigh vacuum (UHV, better than 10^{-9} mbar) may be necessary.

2.2 Relevant processes on the substrate surface

The different processes taking place on the substrate surface during thin film growth are summarized in figure 2. If a sufficient supersaturating of is obtained in the gas phase, atoms or molecules arriving on the surface are adsorbed on the surface by physi- or chemisorption. Depending on the sticking coefficient of the material, a certain amount of adatoms might be desorbed from the surface. The adatoms nucleate on the surface or diffuse to energetically favourable lattice sites or surface sites such as steps or kinks. The diffusion process can be described by the hopping of adatoms to neighboring sites with the hopping frequency ν:

$$\nu = \frac{1}{z}\nu_0 \exp(-\Delta E / kT) \tag{3}$$

The attempt frequency ν_0 is determined by the lattice dynamics, z is the number of neighbouring potential wells and ΔE is the energy barrier for hopping of adatoms from one lattice site to the neighbouring potential well.

The formation of highly ordered monolayers is only possible if the surface diffusion is sufficiently high to enable adatom diffusion toward the surface configuration with minimized free energy. For low temperatures and high deposition rates, the growth mode might be kinetically limited, resulting in defect concentrations and surface morphologies strongly differing from the thermodynamically predicted scenarios.

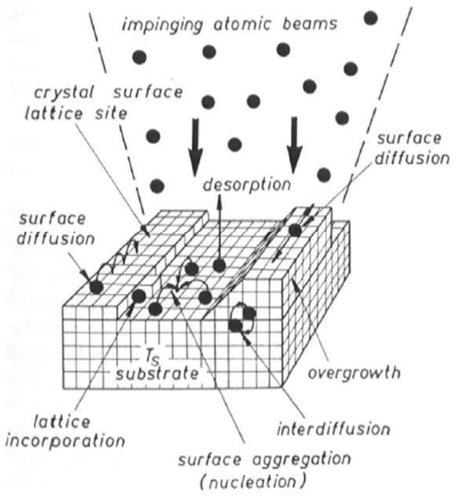


Fig. 2: *Schematic overview over the different processes taking place on the substrate surface[1]*

2.3 Growth Modes

Depending on the growth conditions and the choice of substrate, the microstructure of thin films can vary between amorphous, polycrystalline and single crystalline. Figure 3 depicts the different cases. In the single crystalline case the crystal lattice orientation is identical over the whole thin film. In the case that the lattice is coherent with the single crystal substrate, the thin film is epitaxially grown on the substrate. In specific configurations, thin films can be single crystalline without epitaxial relationship to the substrate.

In polycrystalline thin films, the lattice orientation varies in lateral or even in vertical direction. In amorphous thin films, a long range ordering of the atoms is completely missing.

Epitaxial growth generally requires a substrate with a similar bonding type and a good lattice match. Besides this, sufficient surface mobility of the adatoms is a prerequisite for highly ordered epitaxial growth. In turn, insufficient surface mobility with respect to the timescale of the incoming flux of particles results in the growth of highly defective thin films. In the case of strongly lattice mismatched substrates or distorted substrate surfaces, polycrystalline growth takes place. The size and the texture of crystal grains depends on the specific growth kinetics and can be adjusted by the deposition conditions.

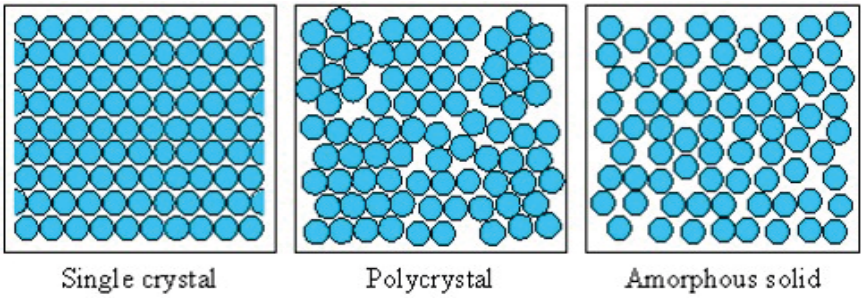


Fig. 3: *Sketch of the different type of thin film microstructure*

Nucleation and growth of a film proceeds from energetically favourable places on a substrate surface and even the cleanest polished surface shows some structure. The characteristic features are the terraces of length, l_s, the steps and the kinks within the step line, which usually runs along well-defined crystallographic directions. If the surface diffusion is fast enough, a randomly deposited adatom will diffuse to the energetically most favourable places like steps and especially kinks. This causes the so called step-flow growth, depicted in figure 4(a), where the nucleation and growth occurs exclusively at steps and kinks. In the case of reduced surface diffusion, e.g. at lower temperatures, several mobile adatoms may encounter each other within a terrace and may form additional immobile adatom clusters within the terraces. By reducing the step distance and hence the diffusion length by vicinal surfaces, the step-flow growth may be extended to lower temperatures.

The details of the growth modes for the simplest case of homoepitaxy, namely, the growth of a film on a single-crystalline surface of the same material, is indicated in Fig. 4. As discussed above, step propagation (figure 4(a)) dominates at higher temperatures and/or small deposition rates and two-dimensional island growth or layer-by-layer growth (figure 4(b))will predominate if immobile clusters are formed by the encounters of mobile adatoms. This simple picture is,

however, quite frequently modified: if the jump across the step is kinetically hindered, multilayer growth will be observed (figure 4(c)). This enlarged activation energy for the jump across the step is called the Ehrlich-Schwoebel effect and can be understood in a simple model as the adatom is nearly dissociated from the surface in the saddle point of this jump.

If we want to grow an epitaxial film on a different substrate (so-called heteroepitaxy), two material parameters have to be considered in addition: the surface energy of the film γ_F, the interface between layer and substrate γ_I, the free surface of the substrate γ_S and the lattice parameter or lattice match of the two materials. In the case of good lattice match, the difference in surface energy leads to two different growth modes as indicated in Fig. 5(a) and Fig. 5(b). As long as :

$$\gamma_F + \gamma_I \leq \gamma_S \tag{4}$$

we observe perfect wetting and pure layer by layer or Frank-van-der-Merve growth. For the opposite case, we observe island or Volmer-Weber growth. For this consideration the surface energies of the crystallographic orientations of actual interest must be applied, which are often not available in data reference tables. If there is a lattice mismatch between substrate and film, an additional growth mode may be observed as indicated in Fig. 5(c) (Stranski-Krastanov growth). A first layer may grow matched to the substrate, which yields additional strain energy. With growing thickness this strain energy increases in proportion to the strained volume and an island formation may become more favourable in spite of the larger surface area.

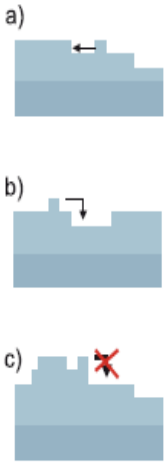


Fig. 4: *Fig. 2: Homoepitaxial growth modes: (a)step-flow growth mode; (b) 2D island growth mode; (c) 3D island growth mode [2]*

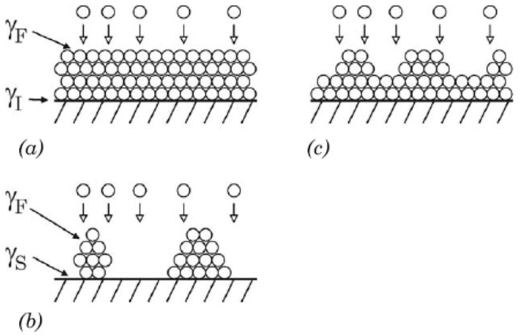


Fig. 5: *Fig. 3: Heteroepitaxial growth modes: (a) layer-by-layer growth, (b) island growth mode, (c) Stranski-Krastanov growth mode*

The contributions of strain and surface energy can quite generally be described in a simple model and the resulting difference in energy between island growth and layer growth is given by equation (5) and its dependence on the island area d is depicted in figure 6.

$$\Delta W = W_{surf} + W_{relax} = const_1 \gamma d^2 + const_2 k \xi^2 d^3 \tag{5}$$

k = bulk modulus, ξ = strain

Considering films of the same volume content, the increased surface energy for the island growth, is proportional to the island area, d^2 , whereas the energy released by relaxation of the lattice is proportional to the island volume, d^3. A relaxation mode which is characteristic of isolated islands is shown in Fig. 7 for a case where the film material has a larger bulk lattice parameter than the substrate. The model predicts a critical value, d_{crit}, where the island growth is finally more favourable and a fast decrease of the energy for larger sizes. However, the limits of the model are reached in this region as the simple relaxation mode is obviously no longer valid for large sizes.

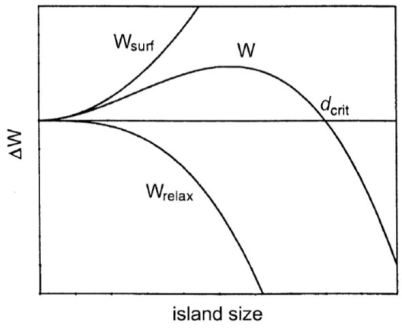


Fig. 6: *Energy contribution as function of the island size according to equation (5)*

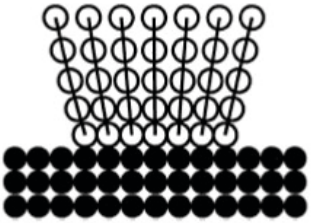


Fig. 7: *Strain relaxation in pseudomorphic (dislocation free) islands [2]*

2.4 Strain relaxation

Along with film growth, the islands will overlap and a closed film will form, which can no longer relax by the mechanism discussed above (Figure 7(d)). A possible mechanism for strain relaxation is the formation of misfit dislocations as schematically shown in Fig. 8. As long as the film is rather thin, there is coherent epitaxial growth on the substrate, however, the unit cell is tetragonally distorted; as the in-plane lattice parameter is forced to smaller values, an expansion, according to Poisson's ratio, is observed in the direction perpendicular to the film. This tetragonal structure is manifested by the different in-plane and out-of-plane lattice parameters and by a tilt of the crystallographic angles. This strain is relaxed by the formation of dislocations as indicated in Fig. 8. and the film returns, in principle, to the cubic structure, however, the interface is only semi-coherent. Since misfit dislocation have a considerable formation energy which has to be overcome by the energy gain of the strain relaxation, a certain critical film thickness has to be reached before dislocation formation becomes energetically favourable. Above this critical thickness, which is determined by the lattice mismatch between substrate and thin film, misfit dislocation are progressively incorporated into the film with increasing film thickness, finally resulting in a relaxation of the thin film lattice to its bulk values.

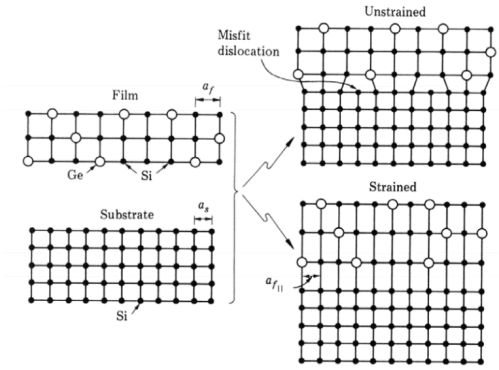


Fig. 8: *Illustration of pseudomorphic strained growth (bottom) versus strain relaxation by the formation of misfit dislocations in a thin film compressively strained by the underlying substrate [3].*

3 Substrates and its surface termination

For the epitaxial and even more for the atomically controlled growth, the importance of the quality of the underlying crystalline substrate cannot be overestimated. For commercial semiconductors (e.g. Si and GaAs) highly perfect single crystals, chemical etching methods to prepare smooth and damage-free surfaces for the epitaxial growth and detailed knowledge about the surface reconstruction all exist and are key to the success of semiconductor technology. For the epitaxial growth of GeTe thin films, Sb-terminated Si(111)-(√3x√3)R30° are advantageous.

For complex oxides, the substrate quality and the related knowledge is in its infancy. Figure 9 depicts an overview over the a-axis lattice constants of commercial available perovkite-type substrates in comparison to the most interesting complex functional oxides. The most common, commercial available substrates for the growth of perovskite thin films are $SrTiO_3$ single crystals, which consists of TiO_2 and SrO sublayers. Usually the as-received substrates have a mixed surface termination.

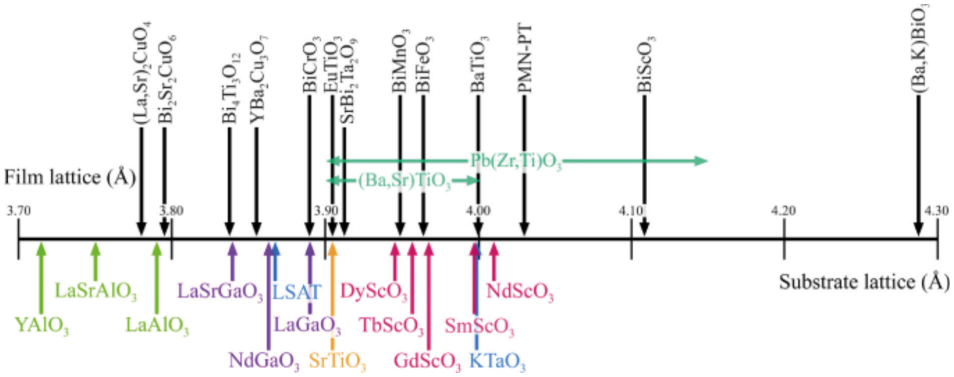


Fig. 7: *Comparison of the lattice constants of several commercial available substrates (bottom) with different complex oxide thin films(top).[7]*

In order to achieve single termination, the SrO has to be removed from the surface by etching the crystal in buffered hydrofluoric acid-solution (BHF). Several recipes and methods are reported in the literature [4], [5].

4 In-situ characterization / RHEED

One of the decisive advantages of MBE is the possibility of implementing all UHV surface analytical techniques and controlling the growth process in situ. Only reflection high energy electron diffraction (RHEED) can be employed at oxygen process pressures of up to 0.5 mbar by adding a partial pumping system (high-pressure RHEED) and can also be implemented into sputter or PLD deposition systems operating at higher pressure. Therefore, we will focus on the process control via RHEED in this chapter.

Diffraction methods are widely used to investigate systems with any kind of translation symmetry, especially crystalline systems. For online control of the growth of nanometer thin epitaxial layers, atomic resolution as well as surface sensibility is required. The resolution of all diffraction methods is roughly given by the wavelength of the incident particles. Thus, atomic resolution can be achieved either by short-wave light such as X-rays or matter waves such as electrons. Since the penetration depth of X-rays exceeds several micrometers it is not a surface sensitive technique so that electrons are utilized. Their de-Broglie wavelength λ_B is given by

$$\lambda_B = \frac{h}{\sqrt{2m_e E_{kin}}} \tag{4}$$

where h is Planck´s constant and me the electron mass.

The kinetic energy E_{kin} is typically in the order of 25keV resulting in a wavelength λB =0.0075nm which is much smaller than the lattice constant of STO. The electron source of RHEED-systems is aligned in grazing incident angle $\alpha \leq 3°$ resulting in two effects: One the one hand, the sample is illuminated along a line across the whole sample so that one gains integrated information all over the sample. On the other hand, as a result of the reduced momentum transfer perpendicular to the surface, penetrations lengths of about 1nm are obtained by RHEED.

A scheme of the experimental configuration can be seen in Fig. 10(a). The incident electrons interact with the sample´s surface and form a diffraction pattern which is visualized by a phosphor screen. If the electrons are scattered by an atomically flat surface, most of them are directly reflected obeying the regular law of reflection resulting in the so-called specular spot.

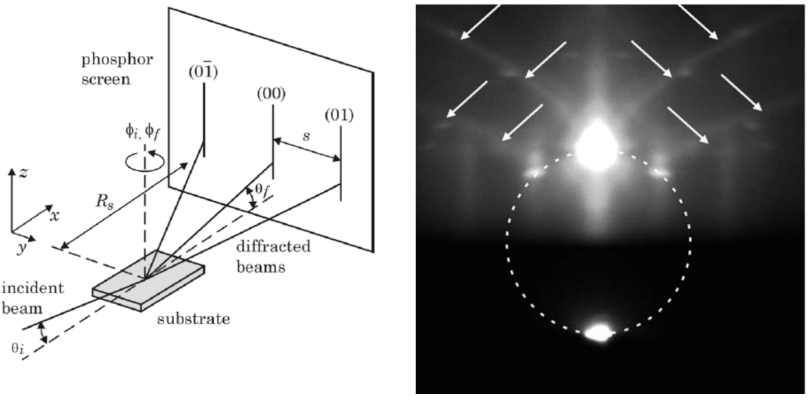


Fig. 10: *(a) Schematic view of the RHEED geometry.(b) typical RHEED pattern of a perfect $SrTiO_3$ surface [6].*

For a perfectly flat surface the maximum number of electrons is regularly reflected. Any roughness, however, causes scatter centres which deflect the electrons in arbitrary directions. For this reason, rough surfaces do not reach the maximum amount of regularly reflected electrons.

Additionally, diffraction maxima may occur besides the specular spot due to in-plane translation symmetry of crystalline samples. Considering a (100)-surface of a cubic lattice, the incident beam can be aligned to the (010) or (001)-direction in order to achieve a perfectly symmetrical condition where the diffraction spots are arranged symmetrically on both sides of the specular spot defining the On-Bragg condition (see Fig. 10(b)). Hereby, all spots, specular and diffracted ones, lie on the so-called Laue circle.

Additionally, Kikuchi-lines, which arise due to multiple scattering effects - including both, inelastic and subsequently elastic scattering - can be observed (see white arrows in Fig. 10(b)). Each line corresponds to a single crystallographic plane of the crystal and can be labelled with a triplet of Miller indices.

When the miscut angle of the substrate is high and consequently the terrace width is short, the specular spot as well as the diffracted spots can split into several smaller spots due to additional diffraction by the terrace structure.

The diffraction pattern in general carries information about the crystallographic structure of the sample since it is mapping its reciprocal lattice. In order to control the growth conditions during the growth process, the analysis of the intensity of the specular spot is of interest because it contains information about the surface roughness as can be shown by the following considerations.

Figure 11 shows the intensity evolution of the RHEED specular spot expected for a perfect layer-by-layer growth mode. Starting in the initial state with an atomically flat substrate, the RHEED-intensity is supposed to be at a maximum. While growing a single monolayer the number of scatter centers is increasing until the coverage reaches 50%. Consequently, the RHEED-intensity is decreasing. When the coverage is larger than 50% the number of scatter centers is decreasing since the holes are filled, i.e. less electrons are irregularly scattered and the RHEED-intensity increases again. After completion of exactly one unit cell, the surface is atomically flat again and the maximum intensity of the specular spot is recovered. As a result, oscillations

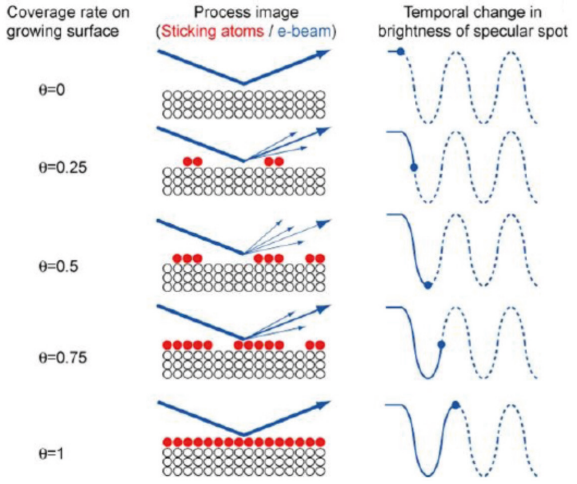


Fig. 11: *Schematic explanation of the diffuse reflection of electrons on the substrate surface during layer-by-layer growth mode and the resulting intensity of the specular RHEED spot [7].*

occur in layer-by-layer growth mode, whereas each maximum corresponds to the growth of a single unit cell.

RHEED-oscillations are used for both, qualitative and quantitative analysis. For MBE growth of complex oxides it is inevitable to employ RHEED oscillations to decide when a single sub-unit layer is completed. In general RHEED enables the analysis of the growth mode and the control of the thin film growth on an atomic scale.

5 Molecular beam epitaxy

5.1 Evaporation techniques

The schematic of the classical MBE source, the Knudsen ore effusion cell, is illustrated in Fig. 12. The evaporation rate, N_e, is described by the Hertz-Knudsen (or Langmuir) equation:

$$N_e = \frac{p_e A_e}{\sqrt{2\pi m k_B T}} \tag{5}$$

p_e is the equilibrium vapour pressure and A_e the area of the aperture [1]. Therefore, the source can be precisely controlled by a single parameter, namely the temperature. However, the technical details are very complex and involve more parameters than shown in Eq. 5 .

Highly reactive materials as well as materials with high melting points can not be evaporated within a Knudsen cell and have to be evaporated by electron beam evaporation. Fig. 13 shows the principle of an electron beam evaporator. The electron beam is magnetically deflected by 270° and is centred on the source material. In this way, a cross-contamination between source material and filament is prevented. On the position where the electron beam hits the source

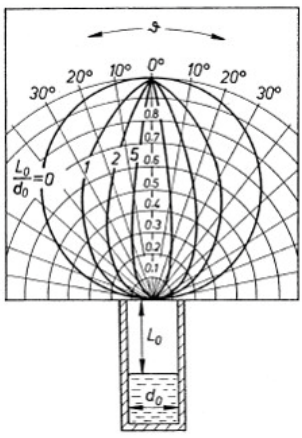


Fig. 12: Schematics of a Knudsen cell and the distribution of the vapour beam intensity [1]. The distribution depends on the ratio L0/d0 and consequently on the filling level of the cell.

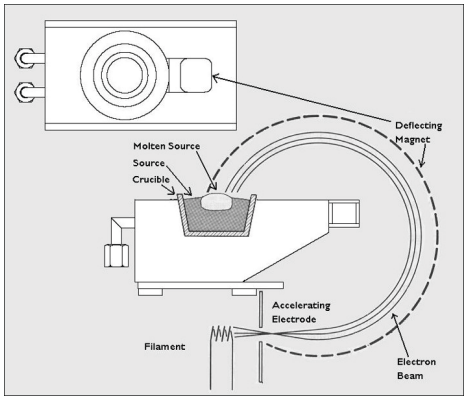


Fig. 13: Schematics of an electron beam evaporator[8]

material, a melt is produced on a block of the same material which can be held in a water-cooled cold crucible in order to avoid contamination of the melt.

It is difficult to evaporate multicomponent materials from a single source, since according to their phase diagram many materials decompose before they evaporate. Since the evaporation rate of a material depends on its vacuum pressure, compounds consisting of elements with strongly differing vacuum pressure will not be transferred to the thin film in a stoichiometric way. In particular for oxides, evaporation in vacuum likely results in the loss of oxygen and in the formation of suboxides. Therefore, oxides have to be evaporated in the presence of oxygen gas.

5.2 MBE systems

MBE offers the possibility to control the stoichiometry of multi-element compounds in a very precise way, by employing separate sources for each element. MBE has evolved from simple thermal evaporation techniques by the application of UHV techniques to avoid disturbances by residual gases. A schematic view and a photo of a MBE system is shown in Fig. 14. The main components are multiple beam sources with shutters, which are crucial for precisely controlling the growth of multi-element materials or multilayers. The substrate can be heated in order to guarantee a sufficient surface mobility of the adatoms. Due to the UHV environment, all UHV surface techniques might be applied for process control. However, RHEED control is inevitable for monitoring layer-by-layer growth and prerequisite of a MBE system. For the deposition of oxides, oxygen gas or ozone is introduced via a leak valve (reactive MBE, or Oxy-MBE).

MBE is an extremely powerful tool for the growth of well defined thin film heterostructures and superlattices. In addition, unlike single source sputtering and laser ablation, MBE does not require the fabrication of the desired compound. Thus, the growth of metastable compounds and structures that cannot be realized by bulk synthesis techniques can be achieved in a MBE system.

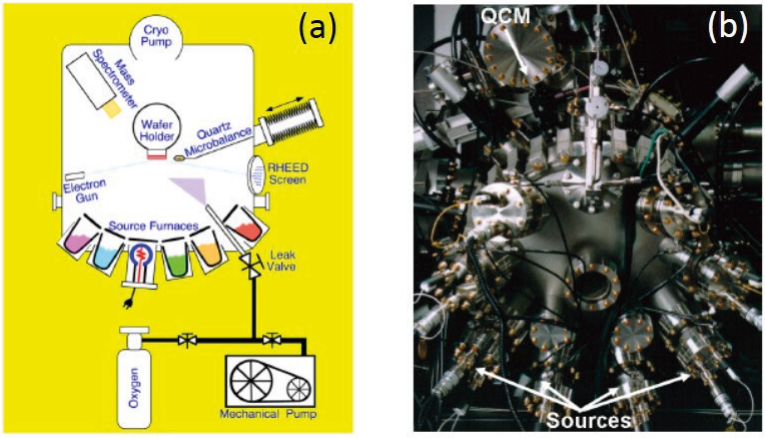


Fig. 14: *(a) Sketch of an Oxy-MBE system (b) Picture of the MBE system at Pennsylvania State University [9]*

For MBE growth of oxides, it is a challenging task to sufficiently oxidize the metallic components by the supply of oxygen gas and at the same time preserving the molecular beam on its way from the source to the substrate. As can be seen from the strong decrease of the mean-free path of different metals (figure 15) with increasing oxygen pressure, the MBE regime is restricted to a pressure of 10^{-6} Torr. On the other hand the stability line of many metal oxides requires an oxygen pressure at the limit of the MBE regime (see figure 16). This problem can be overcome by using an ozone source activated oxygen produced by an rf-source instead of molecular oxygen, which shifts the oxidation of metal ions to lower pressure.

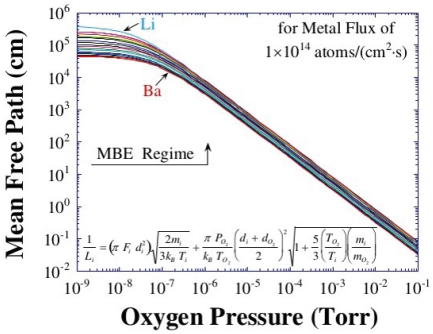


Fig. 15: *Mean free path versus oxygen pressure for a flux of different metals[10]*

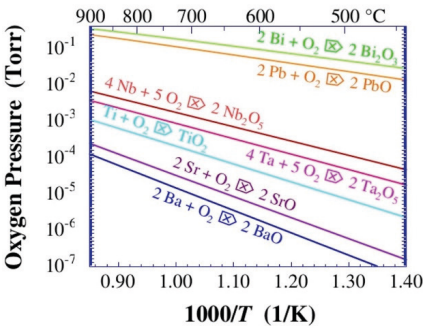


Fig. 16: *Pressure-temperature stability line of different metal oxides[10]*

5.3 Adjusting the stoichiometry in oxide thin films

Reactive MBE can be employed to growth oxide thin films and heterostructures with high crystalline perfection and with precise control over the cation as well as anion stoichiometry. It could for example be demonstrated that the cation stoichiometry of homoepitaxially grown $SrTiO_3$ can be varied systematically from Sr rich over stoichiometric to Ti rich by adjusting the flux from the Ti and the Sr sources, respectively [11]. Furthermore, even artificially designed phases such as the homologous Ruddlesden-Popper series $Sr_{n+1}Ti_nO_{3n+1}$ have been grown by reactive MBE [12]. Figure 17 depicts the high resolution transmission electron microscope images of thin film with n=1-5.

Since oxygen vacancies are of key relevance for the resistive switching properties of oxide thin films, it is highly interesting to adjust their concentration during thin film growth in a controlled way. It has been demonstrated that it is possible to adjust the amount of oxygen vacancies in resistive switching HfO_{2-x} thin films to such a level that the electrical conductivity can be manipulated within several orders of magnitude by changing the oxygen flow rate. Figure 18 shows that the room temperature resistivity changes over several orders of magnitude with the oxygen flow [13]. As a result of the related change of optical band-gap, the thin films change their colour from transparent HfO_2 (inset Fig. 18(a)) to black with golden shine for 0.3 sccm flow rate (inset Fig. 18(c)). Furthermore, it could be demonstrated that the forming voltage of HfO_{2-x} thin film devices can be systematically varied with the oxygen flow rate [14].

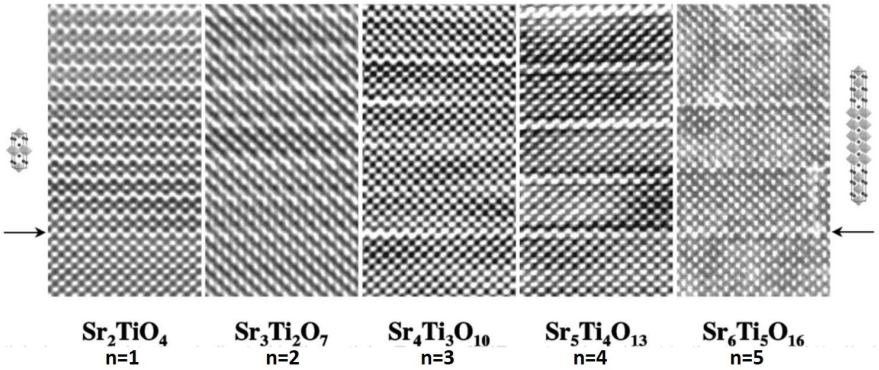


Fig. 17: High resolution transmission electron microscope images of $Sr_{n+1}Ti_nO_{3n+1}$ thin films with n=1-5[12]

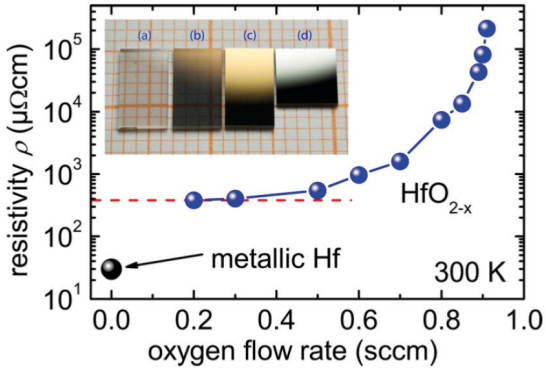


Fig. 18: *Room-temperature resistivity of HfO_{2-x} thin films as function of the oxygen flow rate [13]. Inset: Photo of the thin films produced with different oxygen flow.*

5.4 Molecular beam epitaxy of GST

MBE growth has been employed to grow epitaxial GST thin films on single crystalline slightly mismatched (1%) GaSb (001) and GaSb(11) substrates and single crystalline highly mismatched (11%) Si (111) substrates[15]. Although MBE allows a precise control of the elemental flux, the stoichiometry control of GST is hindered by a strong temperature dependence of the desorbtion of GeTe molecules. Therefore, slight temperature variations during MBE growth might result in composition fluctuation in the thin films [16].

One of the key control parameters for the GST microstructure is the substrate surface. For GeTe thin films grown on Si (111) it has been shown that the domain structure and the resulting surface roughness can be engineered with the surface termination[17].

On Sb passivated Si(111) surfaces, high quality GeTe-Sb2Te3 superlattices could be obtained as shown in figure 19 [18].

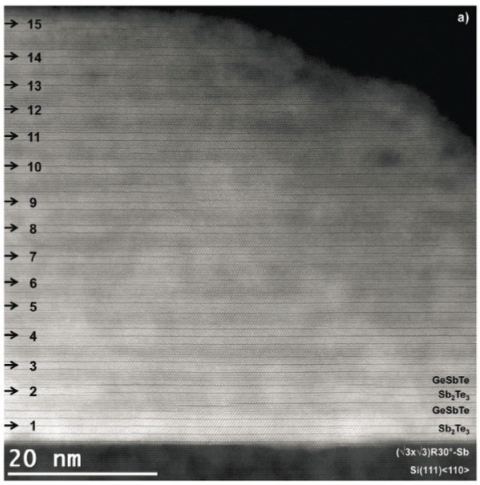


Fig. 19: *Annular dark field image of GeTe(1nm-Sb2Te3(3nm) superlattices grown by MBE [18]*

6 Sputtering

6.1 Basic principle

Sputtering is a popular method for the growth of phase change materials as well as for binary and complex oxides. The sputtering process is schematically illustrated in Fig.20(a). A noble gas ion, usually Ar, which has been accelerated within the darkroom with nearly the full voltage applied of 50 to 1000 V, hits the surface atoms. The following collision cascade leads to back-

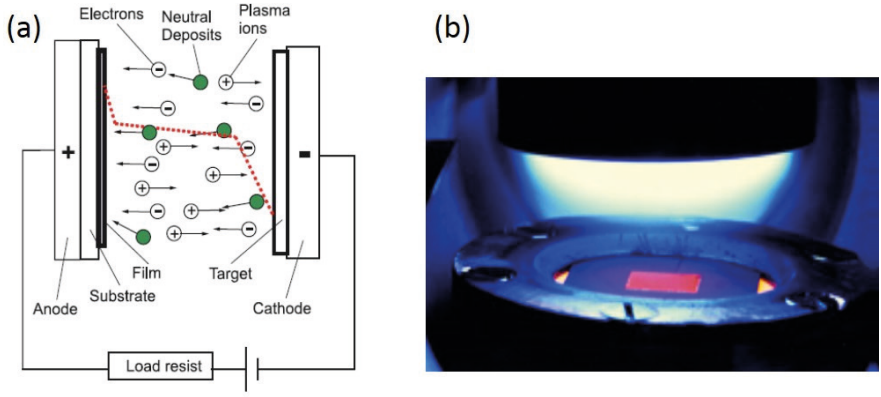


Fig. 20: *(a) Schematics of a DC sputter system; the red dotted line indicates the potential between anode and cathode (b) High pressure DC sputtering deposition at the group of Poppe PGI [2]*

reflected atoms which can leave the surface. The sputtering yield depends on the relative masses of projectile and target atoms. The threshold energy for sputtering is much higher than the surface binding energy, W_b, of the atoms which is of the order of 4 to 8 eV. This difference can be directly understood as several collisions are necessary in order to obtain an atom in the backward direction. Hence, the threshold is observed at $4W_b$ to $8W_b$ resulting in a threshold energy of 20 to 50 eV. A linear increase of the sputtering yield is observed for many conditions up to voltages of 1000 eV. At higher energies, the ions penetrate too deep into the target and the yield decreases again. As a result of the high degree of ionization a plasma is formed which can be seen in figure 20(b).

6.2 Sputtering techniques

The simplest sputtering approach is so-called DC sputtering. In a vacuum chamber the target material, which is eroded, is at the cathode side (negative potential), and the substrate for the film is at the opposite anode side. The potential of several 100 volts between these plates leads to the ignition of a plasma discharge for typical pressures of 10^{-1}-10^{-3} mbar and the positively charged ions are accelerated to the target. These accelerated particles sputter off the deposits, which arrive at the substrate mostly as neutral atoms. The discharge is maintained as the accelerated electrons continuously ionize new ions by collisions with the sputter gas. The potential distribution between anode and cathode is indicated by the dotted red line of Fig. 20(a): As the plasma is a good conductor there is no major potential drop in the plasma region, and due to the different mobility of electrons and ions the main voltage drop is observed at the cathode (dark-room). This potential distribution is advantageous as the acceleration of the sputtering gas ions proceeds directly in front of the target and not in a region far off, where the ions would undergo additional collisions and lose their energy on the long path to the target. An ionization degree of less than 1% of the atoms is characteristic of a plasma and consequently a rather low sputter rate.

DC sputtering works very well as long as the target material shows some electrical conductivity. For insulating targets, however, a high-frequency plasma discharge must be applied in order to avoid the accumulation of electric load. A typical frequency of 13.6 MHz is capacitively coupled to the target and there is only a small voltage decay across the electrode. As the electrons are much faster than the ions a negative potential at the electrodes as compared to the plasma potential evolves during each cycle. With a symmetrical arrangement of cathode and anode we would obtain similar re-sputtering rates and no film growth, however, non-symmetries, which yield some bias voltage, are introduced by the coupling of the RF and by differences in the geometry, i.e., different sizes of target and substrate, and especially by the generally applied grounding of the substrate and the deposition chamber. Nevertheless, deposition rates are much lower than for DC sputtering. To improve the ionization rate, magnetic fields can be used which force the electron onto helical paths close to the cathode and yield a much higher ionization probability. This so called magnetron sputtering additionally allows a lower gas pressure, however, it has the disadvantage of a more inhomogeneous target erosion than a simple planar geometry.

6.3 Modification of thin film morphology

The pressure of the sputtering gas is an important process parameter of the sputtering process which must has to be considered and optimized. The pressure controls the free path length of

the atoms and therefore their energy, angular distribution and finally also their incorporation in the film. For metals, so-called zone diagrams for the film growth have been developed [17], which show a systematic influence of gas pressure and temperature on the grain size and texture of polycrystalline thin films (Fig. 21) [19].

Zone 1: Ts < 0.2 Tm: at these low temperatures no bulk diffusion and only very limited surface diffusion is observed which would allow for crystallite rearrangement. Grain structure is composed of fibres whose size and orientation are determined by the initial random nucleation of the grains. Some shielding effects are visible depending on the angle of deposit incidence. This structure extends to higher temperatures, when the ion energy is reduced due to gas collisions. The size of the fibres increases with temperature mainly following the temperature dependence of the nucleation density.

Zone Ts : (0.2-0.3 Tm): In this transition zone, surface diffusion becomes effective and small crystals of energetically unfavourable orientation are eliminated, i.e. a competitive texture is observed.

Zone 2 and 3: Ts > 0.5 Tm: Evolution of morphology and texture is determined by reconstruction, single crystalline columns are formed with increasing diameter, and, finally, texture is determined by the lowest free energy surface of the crystal. The influence of the Ar pressure decreases.

At high temperatures, high quality epitaxial thin film growth as well as atomically flat surface morphology can be obtained by sputter deposition if sufficiently lattice matched substrates are used.

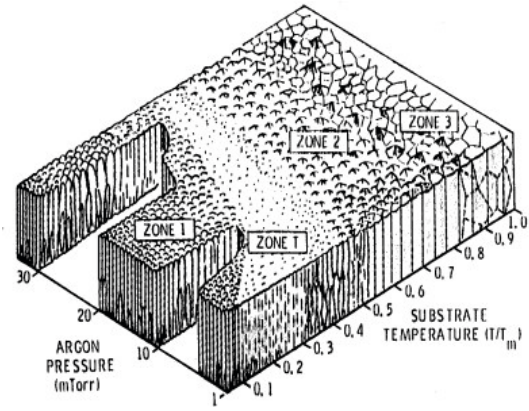


Fig. 21: *Structure zone model for the sputter deposition of metals after Thornton [19]*

6.4 Stoichiometry adjustment during sputter growth

Basically, sputtering from a target containing components with different sputtering yields could result in stoichiometric thin films since the depletion of the target surface with the high-sputtering yield component will quickly balance the excess of this component after a transition period. However, it is often found that sputtering from a single multicomponent oxide target results in non-stoichiometric thin films [20-23]. There exist two mayor reasons for non-stoichiometric transfer during sputtering [24], namely the target surface decomposition during sputtering and the replacement of the target surface depletion by bulk diffusion, which inhibits the self-compensation mechanism of the film stoichiometry mentioned above.

Furthermore, in the case of a high energetic impact of the sputtered species, the bombardment of the growing thin film surface might result in element-selective sputtering or implantation effects causing significant deviations from stoichiometry [25]. In particular, as a result of the admixture of oxygen gas which is needed to grow fully oxidized metal oxides, negatively charged oxygen ions are formed. These are accelerated towards the substrate and result in a considerable bombardment of the growing thin film surface [25, 26]. This problem has been avoided by several approaches, comprising (a) the use of compositionally adjusted targets [20, 21], (b) the use of off-axis geometries [23, 27], (c) the growth in high oxygen gas pressure in order to thermalize the oxygen ions [22] and (d) the precise adjustment of the Ar/O_2 pressure [28]. Therefore, a considerable degree of cation stoichiometry control has been achieved for oxide thin films grown with different approaches of sputtering in the past.

With respect to the oxygen stoichiometry adjustment, it has to be taking into account that the oxygen pressure might also influence the cation stoichiometry by the above mentioned negative ion bombardment or by its impact on the plasma kinetics. This aspect does not play any role for binary oxides, which can be either sputter deposited by a metal target or by a ceramic oxide target. Reactive sputtering from a metallic target, however, leads to metal oxidation which does not exclusively take place within the plasma or the substrate surface but also on the target surface. This so called poisoning of the target surface results in a strong decrease of the deposition rate. In the case of oxygen flow control during sputtering, a change in the metal flux during target poisoning results in an increase of the oxygen partial pressure as a result of the decreased oxygen gettering rate [29]. Pressure control of the reactive gas allows the operation of the sputtering process in the transition regime between elemental and poisoned state of the target and prevents fluctuations of the oxygen partial pressure during sputtering. As a result, an improved precision and run-to-run reliability of the oxygen stoichiometry is obtained with pressure controlled reactive sputtering.

However, oxygen stoichiometry control in typical resistive switching binary metal oxides has been demonstrated also with flow control. In particular, by detailed analysis of the photoelectron spectroscopy, it has been shown for Ta, W and Nb that the oxidation state can be modified systematically with the oxygen flow rate [30](see figure 22).

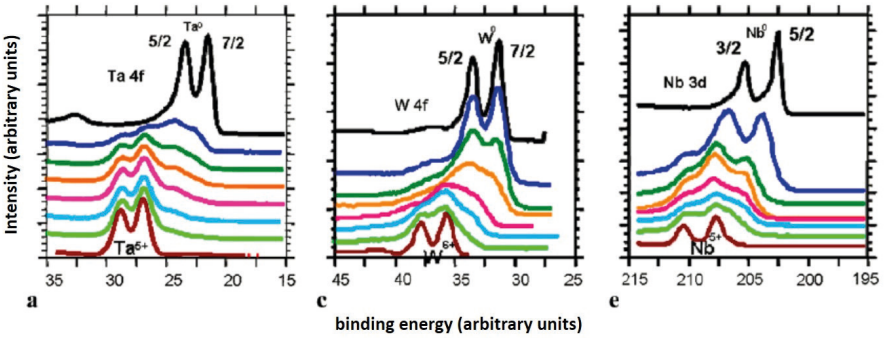


Fig. 22: *XPS spectra of the Ta, W and Nb core levels for thin films sputter deposited with different oxygen gas flow (increasing from top to bottom). [30] The black curves correspond to the metal core levels and the brown curves to the fully oxidized metal core levels.*

7 Pulsed laser deposition

7.1 Basic principle

Pulsed laser deposition (PLD) is employing the laser ablation from a ceramic or single crystal target as particle source. The pulsed laser deposition technique is the most popular method for the deposition of epitaxial complex oxides, whereas due to the presence of highly volatile compounds, PLD did not prevail over the competing other growth method available for phase change materials. A comprehensive description of all details of the PLD technique and the involved physical mechanisms can be found in Ref. [31]. Figure 23 shows a view inside a PLD chamber. A short, ns range, pulse of an excimer laser beam is transferred via a suitable optical system to the rotating target. A satisfactory photon absorption within the oxides is provided by UV-light and therefore usually a wavelength of 248 nm is employed, supplied by a KrF excimer laser. As a result of the short, high energetic laser pulses, the evaporated material is not in the thermodynamic equilibrium and the relative amount of different compounds in the plume corresponds to a certain degree to the target composition even for strongly-differing melting points. The laser energy density on the target is typically 2 – 5 J/cm2 and a crucial point for the deposition of stoichiometric thin films. PLD systems equipped with high-pressure RHEED (Laser-MBE) offer the possibility to engineer thin film growth on an atomic scale.

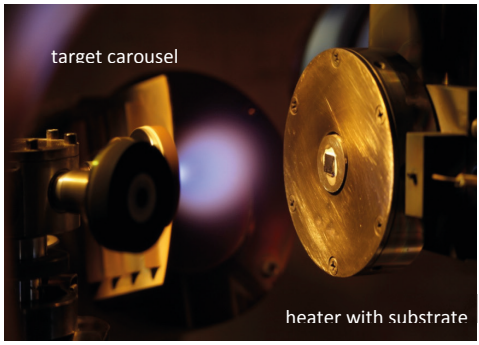


Fig. 23: *Picture of the laser plume within the PLD chamber*

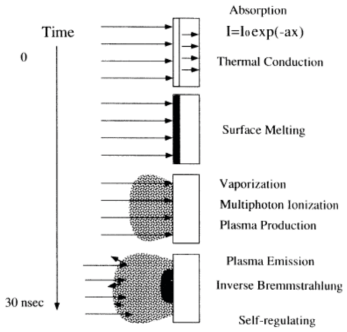


Fig. 24: *Schematic of the time evolution of the laser plume [1]*

The chronological evolution of the plume is depicted in Fig. 25. At the early stage of the laser pulse a dense layer of vapour is formed in front of the target. Energy absorption during the remainder of the laser pulse causes the increase of both pressure and temperature of the vapour and results in partial ionization. This layer expands from the target surface due to the high pressure and forms the so-called plasma plume. During this expansion, internal thermal and ionization energies are converted into kinetic energy of the ablated particles (up to hundred eV).

Attenuation of the kinetic energy occurs during expansion into the background gas due to multiple collisions. Oxides can be deposited in an oxygen pressure of 10^{-6} mbar to 1 mbar. At high pressures, thermalization occurs at a penetration length comparable to the target-to-substrate distance. As a consequence of the attenuation, the kinetic energy of the particles can be varied over an extremely wide range by the oxygen pressure.

In combination with RHEED analysis, PLD offers the possibility for an atomically controlled growth of complex oxide thin films [6]. Complex oxide thin films and superlattices have been grown by PLD with a crystalline perfection indistinguishable from MBE grown thin films.

On the other hand, by tuning the PLD parameters such as temperature and laser fluence, extended defects can be intentionally introduced and employed to improve the resistive switching performance [32], [33], [34].

7.2 Stoichiometry adjustment during PLD growth

It is generally assumed that the cation stoichiometry of the ceramic or single crystalline target is conserved in the thin film. However, this is according to the current knowledge true only to a certain extent. Manifold physical processes determine the cation and anion stoichiometry of complex oxide thin films during PLD growth. The sources of cation non-stoichiometry can be roughly divided into three stages, namely the ablation of material from the target surface, the transport of material in the plume and the nucleation and growth of a thin film on the substrate surface (see figure 25).

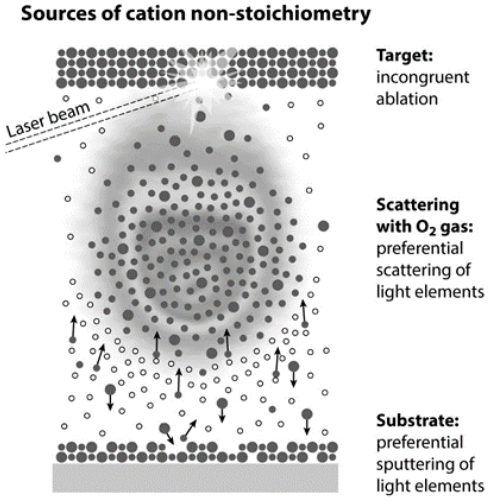


Fig. 25: *Simplified sketch of the different scattering and sputtering processes which could induce cation non-stoichiometry during the different PLD stages. Heavy ions are depicted as large circles and light ions as small circles. Oxygen is depicted as open circle. In order to draw the attention to the cation site, a binary target material with two different metal ions is sketched without showing the oxygen anions in target and thin film* [35].

In the first stage of PLD, the laser light is absorbed in the solid target material and atoms as well as ions and electrons are ejected from the surface. However, the released energy could also be transferred to phonons thereby heating up the solid, finally resulting in a thermal evaporation process. In this case, an incongruent melting of the target material as well as strong differences in the melting points of its components result in non-stoichiometric material transfer. As thermal effects usually dominate in the low fluence regime, a certain threshold fluence is needed to enable non-thermal ablation. If the length scale of the laser pulse is much shorter than the typical time constants for the flow of heat, thermal effects can be neglected. As a rule of thumb for stoichiometric ablation, the volume of the ablated material should be much thicker than the heated volume.

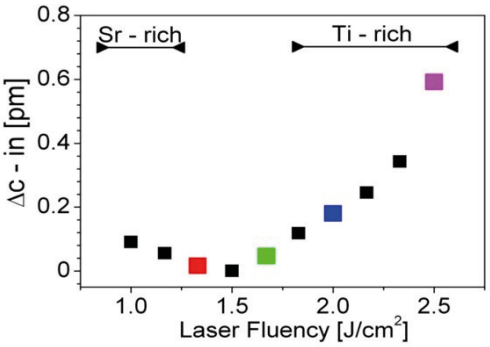


Fig. 26: *Lattice expansion with respect to the bulk $SrTiO_3$ lattice for thin $SrTiO_3$ films in dependence of the laser fluence [36]*

The cation non-stoichiometry of $YBa_2Cu_3O_7$ thin films ablated at low fluence by thermal processes was attributed to the non-congruent melting of YBCO which exhibits a complex phase diagram [37]. However, also for the congruently melting material STO a non-stoichiometric ablation of the cation sublattice has been reported for low laser fluence [37]. In particular it was observed that the Sr/Ti ratio of the ablation spot increases with increasing laser fluence resulting in a considerable non-stoichiometry of the PLD grown $SrTiO_3$ thin films [38, 39].

The homoepitaxial growth of $SrTiO_3$ is the most intensively studied system in terms of the stoichiometry variation with the PLD conditions. One important structural parameter of thin films is the c-axis lattic constant c which can be determined from X-ray diffraction analysis. For the homoepitaxial case, one would expect identical lattice constants for substrate and film. However, Figure 26 shows Δc, the deviation of the c-axis of the $SrTiO_3$ thin film from the $SrTiO_3$ bulk value as a function of the laser fluence F. The X-ray photoelectron spectroscopy (XPS) analysis of the films revealed that $SrTiO_3$ thin films exhibiting a c-axis expansion exhibit a Sr/Ti ratio $\neq 1$ and that films which show no c-axis expansion have a Sr/Ti ratio of ~1 ([40, 41]). Therefore, Δc can be attributed to the non-stoichiometry in a direct way. It could be clarified that the observed variation of the $SrTiO_3$ thin film cation stoichiometry with laser fluence results from a combination between preferential scattering of light plume species, i.e. Ti in $SrTiO_3$, during the propagation towards the substrate and incongruent ablation of the $SrTiO_3$. On the one hand, the Ti-rich films obtained in the high fluence regime result from the increase of Ti-content in the plume due to the preferential ablation of the Ti species with increasing laser fluence. On the other hand, the Sr-rich films obtained in the low fluence regime result rather from the loss of Ti-species during the transfer of the plume from the target to the substrate than from a Sr rich ablation.

These findings for the model system of homoepitaxially grown $SrTiO_3$ can in general be transferred to the growth of other complex oxides containing elements with significant difference in atomic weight and have already been confirmed for other perovskite thin films, e.g. $BaTiO_3$ [42], $La_xBa_{1-x}MnO_3$ [43], $EuTiO_3$ [44, 45] and for $SrTiO_3$ grown on lattice mismatched substrates [46].

Keeping in mind the previously discussed processes, a compound consisting of heavy metals and the extremely light and volatile oxygen will result in a strong oxygen deficiency. In particular, the metal-oxygen bonds are to a large extend either directly dissociated during the laser light-solid reaction on the target surface or afterwards by collisions during plume propagation. However, in contrast to the cation case, the loss of oxygen can be compensated by the background gas during PLD or post-annealing. Furthermore, it has been reported that $SrTiO_3$ films

grown at low pressure are oxidized above temperatures of about 600°C by the underlying $SrTiO_3$ substrate [47, 48].

In thermal equilibrium with the surrounding oxygen environment, the concentration of oxygen vacancies can be nicely predicted at a given oxygen pressure if the defect equilibria for oxygen vacancies are known for the related materials. For PLD growth, thermodynamic considerations have to be handled with care since, as explained in previous sections, the oxygen pressure strongly influences the plume kinetics and thereby the scatting processes taking place in the plume and on the substrate surface. Furthermore, a high amount of oxygen vacancies can be formed in the different stages summarized in figure 25, but might not be sufficiently recovered by the ambient oxygen due to kinetic limitations. It could be demonstrated by quenching experiments that the oxygen vacancy concentration created in thin film and substrate during PLD growth exceeds several orders of magnitude the equilibrium value at a given oxygen pressure and temperature [49].

Besides this, it has been observed [50] that the oxidation of an as-grown STO monolayer is not completed between two laser pulses and may take a longer time to complete than the recrystallization of the monolayer. Based on this phenomenon, it has been proposed that oxygen vacancies can be frozen in the oxide thin films by the adjustment of crystallization and oxidation kinetics [51, 52]. Based on this effect, the oxygen vacancy concentration in TiO_2 thin films has been adjusted by the PLD growth rate [52]. Figure 27 shows that the resistivity and the carrier concentration depends strongly on the growth rate.

Nevertheless, the oxygen vacancy concentration can also be systematically varied with the oxygen pressure during growth, however, one should keep in mind that a change of the oxygen pressure might considerably change the cation stoichiometry as shown for $SrTiO_3$ thin films [35],[53].

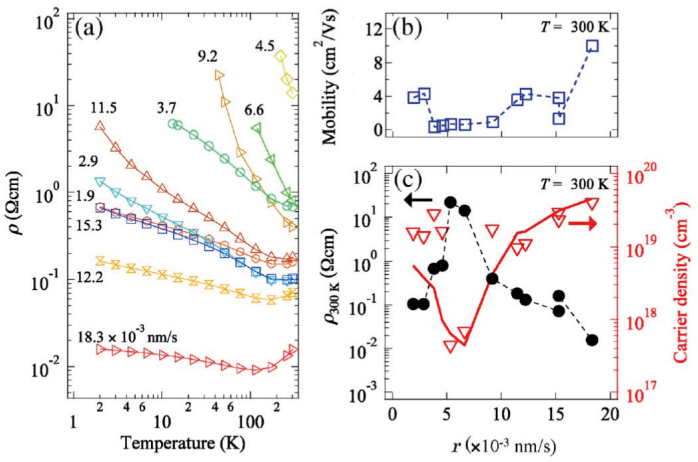


Fig. 27: *Dependence of the charge carrier transport in rutile TiO_2 thin films on the PLD growth rate [52].*

8 Summary

Physical vapour deposition techniques have been successfully employed for the fabrication of memristive thin films. MBE growth offers the highest degree of perfection with respect to stoichiometry control and crystallinity of metal oxides as well as higher chalcogenides. However, MBE requires costly, sophisticated equipment and a high level of expertise for its operation. Sputter deposition is a simple technique, which is widely employed for the growth of binary oxides and GeSbTe alloys in academics and industry. PLD has become the most popular technique for the growth of epitaxial complex oxides. To some extend a stoichiometric transfer from target to thin film occurs by self-adjusting mechanisms for both, PLD and sputtering. However, for both techniques, the plasma kinetics as well as the interaction between high energetic particles and the substrate have to be considered if a high degree of stoichiometric control is required. Both techniques deposit thin films far from thermal equilibrium and kinetic limitations play a key role with respect to thin film morphology as well as defect density. In particular, oxygen vacancy concentrations in oxide thin films might deviate several orders of magnitude from the thermodynamically expected values. For all three physical vapour deposition techniques, a variety of approaches to engineer oxygen vacancy concentrations have been shown which is of key relevance for tuning the resistive switching properties of thin film devices.

References

[1] Hermann, M. A. & Sitter, H. (1996) Molecular Beam Epitaxy, Springer Verlag.

[2] Waser, R. (2003) Nanoelectronics and Information Technology, Wiley - VCH.

[3] Bean, J. (1985) Science, 230 (4722), 127-131.

[4] Kawasaki, M. *et al.* (1994) Science, 266 (5190), 1540-1542.

[5] Koster, G., Kropman, B. L., Rijnders, G. J. H. M., Blank, D. H. A. & Rogalla, H. (1998) Applied Physics Letters, 73 (20), 2920-2922.

[6] Rijnders, G. (2001) PhD Thesis

[7] http:/www.pascal-co.jp

[8] http://www.cleanroom.byu.edu/metal.phtml

[9] Rodriguez Contreras, J. (2003) PhD Thesis Universität zu Köln

[10] Schlom, D.G. &, J.S. Harris Jr. (1995) MBE growth in high Tc superconductors in Molecular beam epitaxy: Application to key materials, William Andrew Pub.

[11] Brooks, C. M., Kourkoutis, L. F., Heeg, T., Schubert, J., Muller, D. A. & Schlom, D. G. (2009) Applied Physics Letters, 94 (16), 162905/1-.

[12] Haeni, J. *et al.* (2001) Applied Physics Letters, 78 (21), 3292-3294.

[13] Hildebrandt, E., Kurian, J., Mueller, M. M., Schroeder, T., Kleebe, H. J. & Alff, L. (2011) Applied Physics Letters, 99 (11), 112902/1-3.

[14] Sharath, S. U. *et al.* (2014) Applied Physics Letters, 104, 063502.

[15] Rodenbach, P. *et al.* (2012) Physica Status Solidi-Rapid Research Letters, 6 (11), 415-417.

[16] Perumal, K., Braun, W., Riechert, H. & Calarco, R. (2014) Journal of Crystal Growth, 396, 50-53.

[17] Wang, R. *et al.* (2014) Journal of Physical Chemistry C, 118 (51), 29724-29730.

[18] Momand, J., Wang, R., Boschker, J. E., Verheijen, M. A., Calarco, R. & Kooi, B. J. (2015) Nanoscale, 7 (45)

[19] Thorton, J. (1974) Journal of Vacuum Science & Technology, 11 (4), 666-670.

[20] Wang, X. *et al.* (1999) Journal of Vacuum Science and Technology A, 17 (2), 564-570.

[21] Tanabe, K., Lathrop, D., Russek, S. & Buhrman, R. (1989) Journal of Applied Physics, 66 (7), 3148-3153.

[22] Poppe, U. *et al.* (1988) Solid State Communications, 66 (6), 661-665.

[23] Eom, C. *et al.* (1989) Applied Physics Letters, 55 (6), 595-597.

[24] K Wehner, G. Betz G. (1983) Sputtering by Particle Bombardment, Springer, Berlin.

[25] Ohring, M. (2002) Materials science of thin films, Academic Press, San Diego.

[26] Shah, S. & Carcia, P. (1987) Applied Physics Letters, 51 (25), 2146-2148.

[27] Lecoeur, P., Mercey, B. & Murray, H. (1995) Journal of Vacuum Science and Technology A, 13 (4), 2221-2227.

[28] Im, J., Streiffer, S. K., Auciello, O. & Krauss, A. R. (2000) Applied Physics Letters, 77 (16), 2593-2595.

[29] Sproul, W.D., Christie, D.J. & Carter, D.C. (2005) Thin Solid Films, 491, 1-17.

[30] Goldfarb, I. *et al.* (2012) Applied Physics A: Materials Science and Processing, 107 (1), 1-11.

[31] Chrisey, D. B. & Hubler, G. K. (1994) Pulsed Laser Deposition of Thin Films, John Wiley & Sons, Inc.

[32] Muenstermann, R. *et al.* (2010) Journal of Applied Physics, 108 (12), 124504/1-8.

[33] Shibuya, K., Dittmann, R., Mi, S. & Waser, R. (2010) Advanced Materials, 22 (3), 411-414.

[34] Raab, N., Bäumer, C. & Dittmann, R. (2015) AIP Advances, 5, 047150.

[35] Dittmann, R. (2015) "Stoichiometry in epitaxial oxide thin films", Epitaxial growth of complex oxides, Elsevier, Cambridge (2015), Elsevier, Cambridge.

[36] Sebastian Wicklein, (2013) Dissertation

[37] Dam, B., Rector, J.H., Johansson, J., Kars, S. & Griessen, R. (1996) Applied Surface Science, 96-98, 679-684.

[38] Dam, B., Rector, J. H., Johansson, J., Hujibregtse, J. & De Groot, D.G. (1998) Journal of Applied Physics, 83 (6), 3386 - 3389.

[39] Wicklein, S. *et al.* (2012) Applied Physics Letters, 101 (13), 131601/1-5.

[40] Ohnishi, T., Lippmaa, M., Yamamoto, T., Meguro, S. & Koinuma, H. (2005) Applied Physics Letters, 87 (24), 241919/1-3.

[41] Ohnishi, T., Shibuya, K., Yamamoto, T. & Lippmaa, M. (2008) Journal of Applied Physics, 103 (10), 103703/1-6.

[42] Kan, D. & Shimakawa, Y. (2011) Applied Physics Letters, 99 (8), 81907/1-3.

[43] Amoruso, S. *et al.* (2010) Journal of Applied Physics, 108 (4), 43302/1-.

[44] Shkabko, A. *et al.* (2013) Apl Materials, 1 (5), 52111/1-9.

[45] Orgiani, P., Ciancio, R., Galdi, A., Amoruso, S. & Maritato, L. (2010) Applied Physics Letters, 96 (3), 32501/1-3.

[46] Breckenfeld, E., Wilson, R., Karthik, J., Damodaran, A. R., Cahill, D. G. & Martin, L W. (2012) Chemistry of Materials, 24 (2), 331-337.

[47] Chen, F. *et al.* (2002) Applied Physics Letters, 80 (16), 2889-2891.

[48] Schneider, C. W. *et al.* (2010) Applied Physics Letters, 97 (19), 192107.

[49] Xu, C. (2015)

[50] Zhu, X., Si, W., Xi, X. & Jiang, Q. (2001) Applied Physics Letters, 78 (4), 460-462.

[51] Ohtomo, A. & Hwang, H. Y. (2007) Journal of Applied Physics, 102 (8), 83704/1-6.

[52] Tachikawa, T. *et al.* (2012) Applied Physics Letters, 101 (2), 22104/1-.

[53] Xu, C., Wicklein, S., Sambri, A., Amoruso, S., Moors, M. & Dittmann, R. (2014) Journal of Physics D: Applied Physics, 47 (3), 34009/1-.

B 3 Nanotechnological Integration

J. Moers
Helmholtz Nanoelectronic Facility
Peter Grünberg Institut, PGI 8-HNF
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Sample Modification **3**
2.1 Oxidation 3
2.2 Rapid Thermal Processing 5
2.3 Ion Implantation 6
2.4 Planarization 8

3 Structure Definition **9**
3.1 Optical Lithography 10
3.2 Exposure wavelength and light sources 16
3.3 EUV-Lithography 17
3.4 E-Beam Lithography 19
3.5 NIL 21
3.6 Overlay 23

4 Structure Transfer **24**
4.1 Wet Chemical Etching 25
4.2 Dry Chemical Etch 26

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

In research as well as in production the realization of nanoscaled structures is a basic prerequisite to follow a desired aim. Be that to investigate a new phenomenon at those nanoscaled structures, to realize a prove of concept device in application oriented research or to produce billions of identical devices in an industrial production line. The challenges to be faced are enormous. Critical dimensions dropped down to the Nanometer range as well as in vertical as in lateral orientation. While in vertical direction the thickness in principle can be controlled by controlling the deposition thickness of the material (and in deposition techniques as ALD and MBE this can be done in the monolayer range), in lateral direction methods of lithography have to be employed; which kind depends on how many samples you want to do in one run. As more samples as more challenging this becomes.

To get an impression how challenging this is, let´s do a thought experiment: In semiconductor industry the integrated circuits are processed on silicon wafers with a diameter of 300 mm. If we scale this up by a factor of 1 million, these wafers would be 300 km in diameter, which would cover the whole state of North-Rhine-Westphalia. Printed gate length of a transistor is 22 nm (www.itrs.net), which correspond to 22 mm on the "scaled" wafer (the width of a ruler). The Overlay accuracy in industrial production is 4.5 nm. On the scaled wafer this means that you have to be able to place a ruler with an accuracy of 4.5 mm somewhere in NRW; but here you have to keep in mind, that this accuracy is requested in respect to alignment markers, one of which is in Aachen and the other one in Lemgo (Fig. 1).

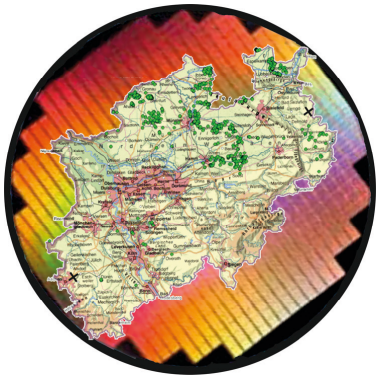


Fig. 1: *Size comparison: 300mm wafer scaled by a factor of 1 million in comparison with the federal state of North Rhine Westphalia (NRW). In this scale nanotechnology means that in industrial production structures in the size of 22 mm are placed with a spatial accuracy of 4.5 mm all over NRW.*

Besides the industrial requirements, where billions of similar devices are produced, in research and development single devices or a limited number of devices are produced. But here structure sizes may be below 10 nm. To cope with these different requirements, other technological solutions for device processing are used: While in industry highly parallel processes are favored, in research flexible and highly serial methods are of advantage. In this chapter we want to give a survey of the different techniques used for nanoscale processing.

The different processes can be classified in three groups: sample modification, structure definition and structure transfer. While in sample modification the whole surface is altered by the process, with the structure transfer methods those structures which are defined with the struc-

ture definition methods are implemented locally on the sample. Typical sample modification processes are oxidation, thermal processing, ion implantation and planarization, structure definition normally is done by lithography processes and sample transfer by etching or local deposition.

2 Sample Modification

2.1 Oxidation

Silicon microelectronics was based on the almost ideal behavior of the interface Si to SiO_2. The SiO_2 can be made with high breakdown voltages and low defect densities, while there are low interface state densities, which is favorable in device applications. Furthermore insulation layers of SiO_2 can easily be obtained.

The most applied technology is thermal oxidation. Fig. 2 shows a schematic view of a thermal oxidation furnace. The silicon substrates are placed under N_2 atmosphere in a quartz liner tube and heated up to temperatures between 700°C and 1100°C. After heating up, the atmosphere is enriched with an oxidizing agent as O_2 or H_2O. This agent reacts with the silicon forming SiO_2. Oxidations in O_2 are called dry oxidation and in H_2O wet oxidation. O_2 is led into the liner tube from the gas supply, while H_20 is produced by burning H_2 with O_2 to H_2O directly before the chamber. For Hydrogen and Oxygen can be produced with a higher purity as water, this method ensures that the steam is of high purity. For cleaning the furnace HCl gas can be attached to the system.

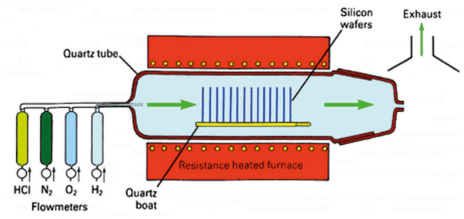


Fig. 2: *Schematic view of an oxidation furnace system. The substrates located in a quartz liner tube are heated to the desired process temperature by a resistance heated system. The gases are led into the reaction chamber by a gas control system. The remaining gases are treated by an exhaust system [36].*

The process was described by Deal and Grove in 1965 [1]. The oxidation of the silicon surface can be characterized into two phases: the initial phase when the oxide is still thin and the diffusion controlled phase. In the initial phase the growth rate is governed by the reaction rate of Si and the oxidation agent, and not by the rate with which the agent arrives at the surface. In the diffusion controlled phase the agent has to diffuse through the already existing SiO_2 layer to react at the Si-SiO_2 interface to new SiO_2. In the initial phase the oxidation rate is constant, hence the resulting oxide thickness d_{ox} is ~ t (where t is the oxidation time). In the diffusion based phase the diffusion of the agent from the gas phase to the surface, from the surface to the interface and the consumption of the agent by the reaction itself has to be taken into account. Doing so the general equation of thermal oxidation can be derived: $d_{ox}^2 + Ad_{ox} = B(t+\tau)$. B is called the Parabolic Rate Constant and B/A the Linear Rate Constant; both vary in temperature as $\exp(-E_A / k_B T)$. E_A is an activation energy and A depends on diffusivity of the agent through SiO_2 and the reaction velocity of the agent with Si. B also depends on the diffusivity and the agent concentration on the gas phase. τ is of relevance

when at $t = 0$ $d_{ox,0}$ is not zero; it is the time which is needed to get $d_{ox,0}$. So the oxide thickness after time t can be described as

$$d_{ox} = \frac{A}{2}\left\{\left(1 + \frac{t+\tau}{A^2/4B}\right)^{1/2} - 1\right\} \tag{1}$$

For short oxidation times where $t + \tau << A^2/4B$ eq.(1) reduces to

$$d_{ox} = \frac{B}{A}(t+\tau) \tag{2}$$

So for small oxidation times and small initial thicknesses the resulting oxide thickness is proportional to the oxidation time (hence B/A is called the Linear Rate Constant). For long oxidation times, where $t >> A^2/4B$ eq. (1) reduces to

$$d_{ox} = \sqrt{Bt} \tag{3}$$

Eq. (3) is called the Parabolic Oxidation Law and this explains why B is called the Parabolic Rate Constant. In this regime, the oxide thickness d_{ox} increases with the square root of time.

Keep in mind that B and B/A dependent on temperature as $\exp(-E_A/k_BT)$ and on the diffusivity. Hence with higher temperatures the oxidation rate increases and it depends on the crystal orientation. This correlation can be seen in Fig. 3.

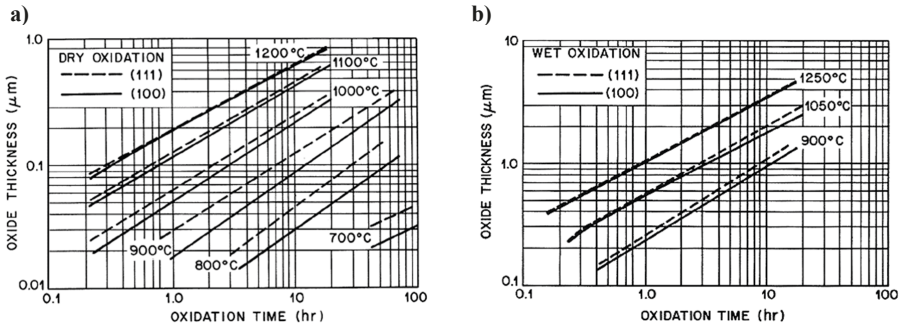


Fig. 3: *Resulting oxide thickness for different temperatures and crystal orientation for a) dry oxidation and d) wet oxidation [36].*

As mentioned, by thermal oxidation SiO_2 layers of high quality can be obtained. The low defect and charge densities of the layer and the low interface state density of the interface to Si enabled a good controllability of a MOSFET channel. Besides this important property, also the insulating properties are widely used. Due to the diffusion controlled process, it is possible to oxidize the Si surface locally by adding a structured diffusion barrier. In this so called LOCOS process a thin oxide layer and a thick Si_3N_4 layer are deposited by CVD, patterned by lithography and etched by reactive ion etching. Hence parts of the surface are masked by the nitride layer and parts are not. During the subsequent oxidation, the uncovered parts are oxidized, while the oxidizing agent cannot diffuse through the nitride, hence oxide cannot grow there. After oxidation, the nitride is removed.

2.2 Rapid Thermal Processing

Besides oxidation processes, thermal treatments of samples are done to activate dopands, anneal layers or interfaces or improve contacts. The major drawback of conventional furnace processes is the high thermal budget the sample suffers from: the thermal treatment will improve certain properties, but it will degrade others, too. One issue is that in conventional processes the whole furnace has to be heated to the desired temperature, increasing the thermal load without an adequate benefit: dopands will diffuse also during the temperature ramping. While structure sizes were scaled down within the last decades, this problem gains impact. To solve this problem the Rapid Thermal Processing (RTP) was developed: in those so called cold wall processes only the sample is heated, not the furnace itself.

Fig. 4 shows a schematic cross section view of an RTP system. The wafer is located on a quartz tray, which is mounted in the middle of a quartz liner tube. Above and underneath the liner tube there are high power lamps. The whole system of liner tube and lamps is embedded in light tight frame. The frame is water could and the liner tube and lamps can be cooled by N_2 flow. Furthermore it is possible to set a defined atmosphere in the liner tube, e.g. you can oxidize the sample by using O_2 or N_2O or you can chose an inert atmosphere using N_2 or Ar. The emitted radiation of the lamps penetrates the quartz without being absorbed and can reach the sample without loss of intensity. The sample itself will absorb the radiation and hence will be heated by the absorbed energy. Fig. 5 shows a comparison between the temperature profiles of an RTP and a conventional process. For only the sample is heated in RTP, temperature ramps of up to 350 K/s are possible. Due to those quick heating and cooling ramps high temperatures can be reached without increasing the thermal stress unnecessary [2].

The measurement of the actual sample temperature T is one of the difficult problems in an RTP system. With the measured T the lamp power is controlled to obey the desired temperature profile. The wafer has a small thermal mass only and is heated rapidly. Using a thermocouple attached to the sample is error-prone due to the quality of the contact between sample and thermocouple. On the other hand, as any other body, a wafer emits heat radiation, which directly depends on T. Therefore the actual temperature of the sample could be determined by so called pyrometer. But the spectrum of the heat radiation is also dependent from the surface of the sample: if there is a poly-Si layer on the surface or not will change the measured temperature (actually: also the thickness of the poly-Si layer will alter the measured temperature [3]). So for each process and for each kind of sample special effort has to be spent to determine the correct temperature and hence enable the power control unit to do the correct temperature profile.

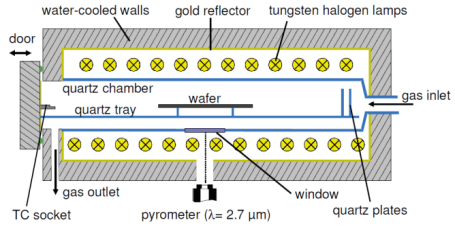


Fig. 4: *Schematic drawing of a cross section through the single-wafer RTP reactor SHS100 [33].*

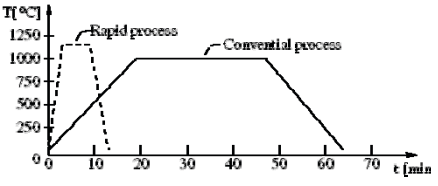


Fig. 5: *Reduction of thermal budget due to quicker temperature ramps [34].*

Rapid Thermal Oxidation

In RTO thin thermal oxides below 10 nm can be obtained easily. Using O_2 as oxidizing agent in the liner tube during process will let a thin SiO_2 layer grow on Si with qualities comparable to conventional furnace processes. Using N_2O will form an oxynitrid which improves break through voltage.

Dopand activation and implantation damage anneal

Doping samples by ion implantation will induce damage on the implanted areas of the sample and the dopands are not on crystal sites and therefore not active electrically. By RTP processes the damages are annealed and during this reconstruction of the crystal lattice the dopands are integrated into the lattice and hence activated.

Contact formation

The formation of ohmic contacts on semiconductor surfaces depends on two issues: doping level and alloying suitable metals with the semiconductor surface. E.g. annealing Ni on Si forms $NiSi_2$, which has low contact resistance to Si.

2.3 Ion Implantation

The most important method to change the electrical properties of given materials is ion implantation. **Fehler! Verweisquelle konnte nicht gefunden werden.** shows a schematic view of an ion implanter: in an ion source the atoms to be implanted are extracted and ionized. A mass spectrometer filters the ions with the correct charge to mass ratio, so that after this filter only the desired species of the element remains. The ion are accelerated to a defined energy and shot on the sample. The impinging ions penetrate the surface of the sample and are decelerated by the electrons and the nuclei of the crystal lattice. The resulting doping level depends on the implantation dose (ions/cm^2) and the resulting implantation depth.

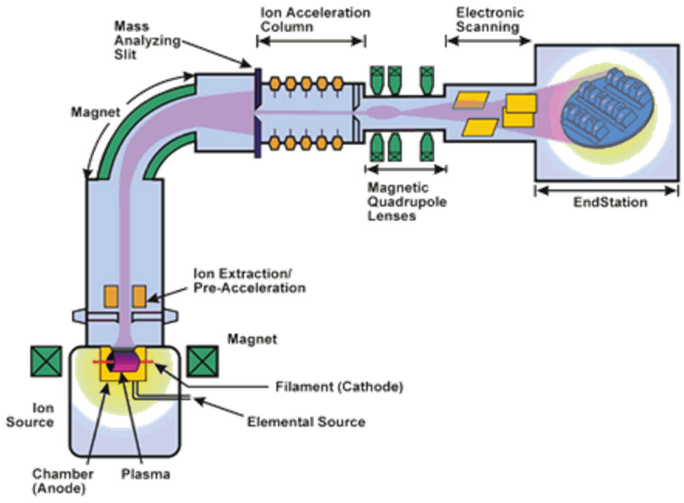


Fig. 6: *Schematic view of a ion implanter [4].*

The dose D can be calculated by the ion current I, the implantation time t and the implantation diameter A. The dose is given as

$$D = \frac{I \cdot t}{e \cdot A} \tag{4}$$

Given a current I of 1 mA and a dose D of 10^{15}cm^{-2} about t=50 s are needed to implant one 200 mm wafer.

For the deceleration of the ions in the crystal lattice and hence the range distribution of the implanted ions two mechanisms are determining. On the one hand, the inelastic scattering of the ions described by the deceleration cross section $S_e(E)$, where E is the ion energy. On the other hand the elastic scattering on the nuclei in the lattice, describes by the $S_n(E)$. The range distribution of ions in solids can be expressed approximately as [5]:

$$D(x) = \frac{D_0}{\sqrt{2\pi}\Delta R_p} \exp\left[-\frac{(x - R_p)^2}{2\Delta R_p^2} \right] \tag{5}$$

Here R_p is the mean range and ΔR_p the standard deviation. This is a Gaussian shape of the range distribution. Fig. 7 shows the comparison of a SIMS measured implantation profile (dotted line) and a Gaussian fit for an implantation of Boron at 80 kV acceleration voltage and a aerial dose of D_0=10^{15}cm^{-2}. Fig. 8 shows the dependence of R_p on the ion energy.

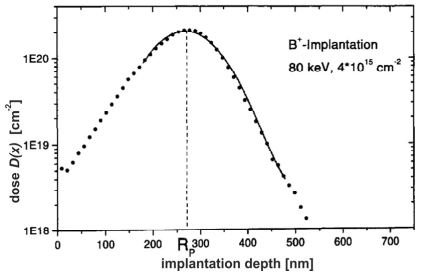


Fig. 7: *Comparison of measured depth profile and Gaussian Fit [6].*

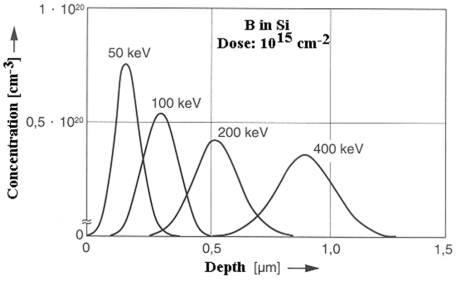


Fig. 8: *Implantation depth in dependence of ion energy [32].*

All this considerations are for amorphous systems. In crystalline substrates the effect of channeling occurs. Due to the high ordering of the atoms in a crystal lattice, the projected surface density of atoms is strongly dependent on the orientation. When looking on a (100) silicon surface only the upper four atomic layers are visible; further atoms are shadowed by the atoms of the first four layers. Ions not hitting atoms in those four layers will not hit atoms from deeper layers, hence do not dissipate energy. The ions will travel through a channel in the crystal. Therefore the range is enlarged. This effect can be avoided by turning the sample in respected to the incoming ions. Then the surface atoms do not shadow the atoms in deeper crystal layers. An angle of 7° is enough to avoid channeling effectively.

On problem arising during ion implantation is the implantation damage. The stopping ions induce defects into the lattice. Due to their high energy heavy ions will knock crystal atoms from their lattice sites the whole way through the crystal, while light ions do this only at the end of their path (End of Range damage, Fig. 9). Furthermore implanted ions will stop in the

crystal at interstitial sites. To remove those damages, a thermal treatment is applied. During this treatment, the knocked out atoms will integrate themselves on lattice sites again; the same is true for dopand atoms which will become electrically activated by this.

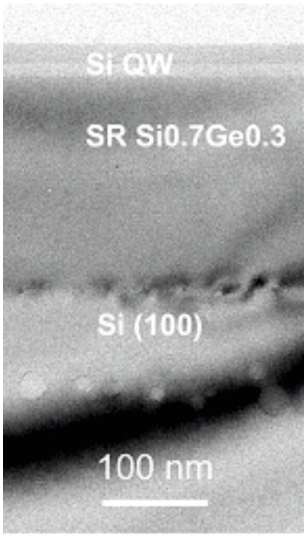


Fig. 9: *TEM cross sectional view of a Si/SiGe/Si layer sequence implanted with H_2. In the substrate layer the end of range damage of the H_2 implantation is clearly visible [38].*

2.4 Planarization

Etching groves or mesas on a sample or adding metal leads will cause topography on the sample surface. Those height differences may cause problems in later device processing or for the functionality of the device. Consider a vertical resonant tunneling diode as described in [27]. A vertical layer sequence is patterned lithographically and etched to columns by RIE. The current flow is perpendicular to the surface, which means, that we have to apply contact at the bottom and at the top of the column. Therefore a thick oxide was deposited and the new surface planarized. In case of [27] the oxide was deposited by spinning on a liquid oxide (HSQ). Spinning on those liquid oxides will result in smooth surfaces which will cover small structures without disturbing the surface. The other possibility is deposition a thick layer e.g. by CVD (cnf. Fig. 10). The deposited layer encloses the structure conformal (Fig. 10b), after planarization step (Fig. 10c) the surface is flat and enables to contact the bottom and the top of the structure (Fig. 10d).

In microelectronics planarization gained impact due to the multilevel metallization: the wiring between the active elements in an IC requires more than one level of metal leads to ensure the performance of the IC. Here a grove in an oxide is filled by depositing a metal (e.g. copper) and the surface layer is polished away. Here the effect arises, that the metal in the middle of the grove is removed quicker that outside the grove. This leads to a bowed surface of the metal in the grove. For typical interconnect applications in an IC this so called dishing effect is independent of structure size but a function of process parameters only [28]. Polishing itself is done by Chemical Mechanical Polishing (CMP). Fig. 11 shows a schematic view of a CMP: a table covered with a polishing pad is rotating with a constant velocity. Onto this pad the sample is pressed with a also rotating wafer carrier head. As abrasive slurry is used spread on

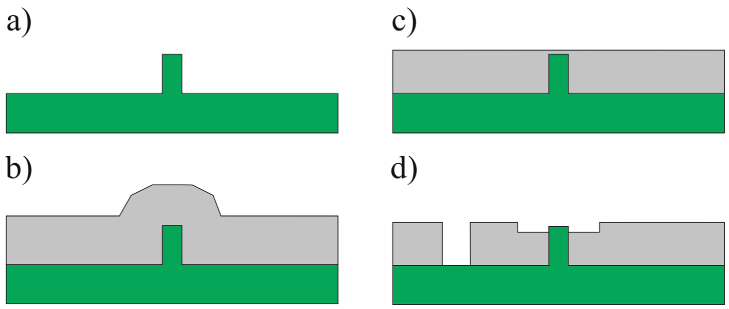


Fig. 10: *Planarization procedure (a) structure to be planarized (b) deposition of thick SiO_2 layer (c) planarization by means of CMP (d) applying vias to perform contacts*

the pad. The homogeneity is dependent on the rotation velocities and the pressure with which the sample is pressed on the pad.

One drawback of the method is that the results depend on the areal density of material to be polished: If there are large structures and large empty areas the process does not distinguish between areas on a structure or empty areas: the material will be removed with the same rate leading to a not planarized surface. This is avoided by adding slots in large structures (area of surrounding material in the structure) or structures in empty areas with no other means than filling those areas.

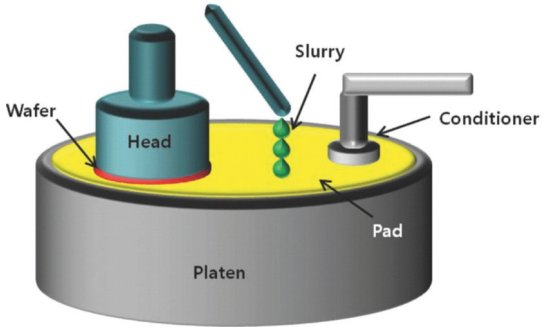


Fig. 11: *Schematic view of a Chemical Mechanical Polishing tool: A table covered with a polishing pad is rotating with a constant velocity. Onto this pad the sample is pressed with an also rotating wafer carrier. As abrasive slurry is used spread on the pad. The homogeneity is dependent on the rotation velocities and the pressure with which the sample is pressed on the pad [37].*

3 Structure Definition

The key process of nanotechnological integration to define lateral structures with dimensions down into the nm-range is lithography. The lithography methods can be categorized into three groups: optical lithography, electron beam lithography (EBL) and nano imprint lithography (NIL). While the first and the second uses radiation (light and electrons, respectively) to transfer the pattern, in NIL a mold is pressed in a deformable polymer and hence the structure of the mold is transferred into the polymer.

While highly advanced optical lithography is used for industrial production (due to its highly parallel processing), in research more elementary kinds of optical lithography are used to define structures with resolutions into the range of several 100 nm. EBL offers a high resolution (≤ 5 nm), but highly serial character of structure definition. Therefore it is not as important for industry, but for research. NIL can be used as an intermediate technology: the area where the structure is defined simultaneously is a few square centimeters, but its resolution is in the order of a few 10 nm.

3.1 Optical Lithography

Optical Lithography is the most important type of lithography. Originally the name referred to lithography using light with wavelength in the visible range. Nevertheless, gradually, the wavelength was driven down to 173 nm, which is used in semiconductor production nowadays, and even shorter wavelengths down to the sub-nm range are under investigation. As the name states light is used to transfer the pattern from a mask to the sample. The light, generated by a suitable light source, is formed to a parallel bundle by a condenser optic. The mask is a transparent (normally quartz) disk coated with an opaque layer, in which the structure to be transferred is inscribed. So the mask is partially transparent to the radiation used. Depending on the illumination scheme used, a shadow of the mask is casted to the wafer, or an image is projected on the samples surface. On the sample there is a so called resist, which is sensitive to the radiation used; the resist is altered by exposure and can be etched away selectively to the not altered resist during development.

When a shadow of the mask is cast to the sample, either the mask is directly in contact with the sample (*contact lithography*, Fig. 12a) or there is a gap *g* between mask and sample (*proximity lithography*, Fig. 12b). Second, an image of the mask can be projected to the wafer by a projection optic (*projection lithography*, Fig. 12c).

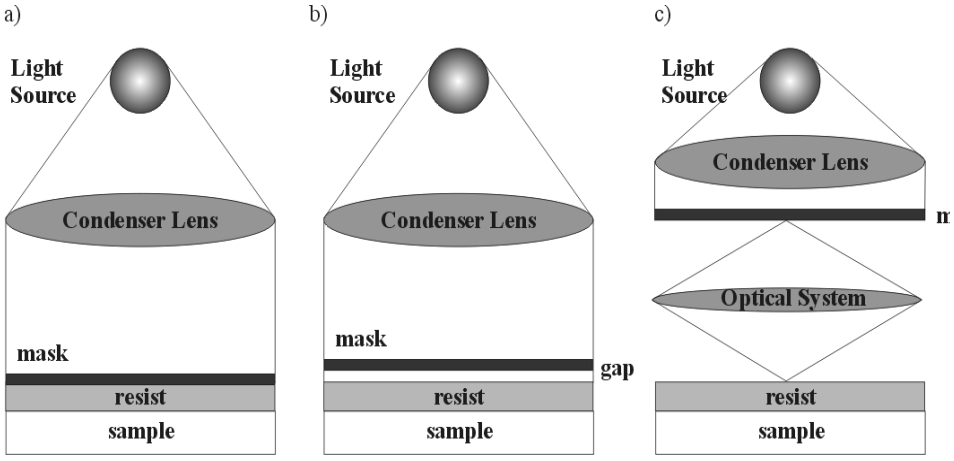


Fig. 12: *Lithography Methods: (a) contact, (b) proximity and (c) projection lithography*

Resolution Limits

The key issue of lithography is the resolution of the system, and hence the size of the smallest feature (Minimum Feature Size: MFS), which can be defined on the sample. This MFS depends on the illumination method, the illumination wavelength λ, on the materials of the optical system and on the resist used.

For contact and proximity lithography the resolution is limited by deflection. For contact lithography $\mathrm{MFS} = \sqrt{d \cdot \lambda}$ yields, where d is thickness of the resist and λ the wavelength. The major drawback for this method is that the quality of the mask is suffering from the contact to the resist, leading to failures in the structure. This problem is avoided using *proximity lithography*. The gap g between sample and mask prevents deterioration of the mask. Drawback is the worse resolution limit given by $\sqrt{(d+g) \cdot \lambda}$.

The method used today in industrial production is the so called *projection lithography*. The mask is not in contact with the sample, so there is no deterioration as in contact lithography, but the resolution is better as in proximity lithography. Furthermore, for an image of the mask is transferred to the wafer, it is possible to reduce this image in size, so the patterns on mask are allowed to be bigger than the patterns on the sample. This is advantageous for the mask fabrication: Errors are also reduced. If it is possible to obtain masks with an accuracy of 100 nm, than the error for a structure of 500 nm to be transferred onto a sample is 20%, if it is transferred one by one. If the image is reduced 4 times, than for a 500nm feature on the sample, the feature on the mask is allowed to be 2µm; therefore the mask error is only 5%. With a reduction optic it is not possible to expose the wafer in one shot. The wafer is located on an x-y-table and is moved ("stepped") under the projection optics and is exposed die by die.

In projection lithography MFS is limited by diffraction. A parallel bundle of light passing a hole P with diameter b will cause a diffraction pattern in distance d behind the slit due to interference between the light passing the slit at the lower rim or in the middle. For two adjacent holes P and Q, the two diffraction patterns overlap. Fig. 13 clarifies the connection between mask (E), diffraction and intensity distribution on the wafer in the image plain (E'). Due to diffraction two sharp features P and Q on the mask give rise to an overall intensity distribution on the sample. To resolve those two features the intensity distribution has to have a minimum between the two main maximums. It is useful to define the so called Modulation Transfer Function (MTF), which is defined as:

$$\mathrm{MTF} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{6}$$

The higher the value - the higher the difference between the maximum and minimum intensity - the better is the contrast between exposed and unexposed areas, the better is the resolution of the equipment. It should be noted, that the MTF is only derived by properties of the optical system. It is a measure, how capable the lithographic tool is in printing structures.

According to the Rayleigh Criterion [7] the two features P and Q can be distinguished, when the maximum of one diffraction pattern is in the first minimum of the second one. In lithography, the Rayleigh Criterion can be expressed as:

$$MFS = k_1 \cdot \frac{\lambda}{NA} \tag{7}$$

where *NA* is the numerical aperture of the optical system given by $NA = sin(\alpha) \cdot n$ (α half the opening angle of optics and *n* index of refraction of medium between last optical element and substrate) and k_1 a constant (typically 0.5-0.9), which accounts for non ideal behaviour of the equipment (e.g. lens errors) and for the influences, which do not come from the optics (resist, resist processing, shape of the imaged structures,...). Therefore k_1 is called the technology constant.

The Rayleigh Criterion is just a first approach to the MFS in microlithography. The mask patterns are not independent (i.e. incoherent) ideal point sources, but they have a finite width and the light is partially coherent. Nevertheless, the form of the criterion gives the right dependencies. If the wavelength is decreased by 10% or the NA is increased by 10%, the MFS is increased by 10%. Therefore by defining the technology constant k_1, the Rayleigh Criterion is the right expression for the MFS in optical proximity lithography.

Another important feature of optical lithography is the Depth of Focus (DOF) which is given by:

$$DOF = k_2 \cdot \frac{n\lambda}{NA^2} \tag{8}$$

Here, k_2 is also constant depending of the technology. A structure, which has to be exposed on a prepatterned substrate, has to be focused over the whole range of the topography of the structure. Therefore, the DOF has to meet the requirements given by the pattern on the sample.

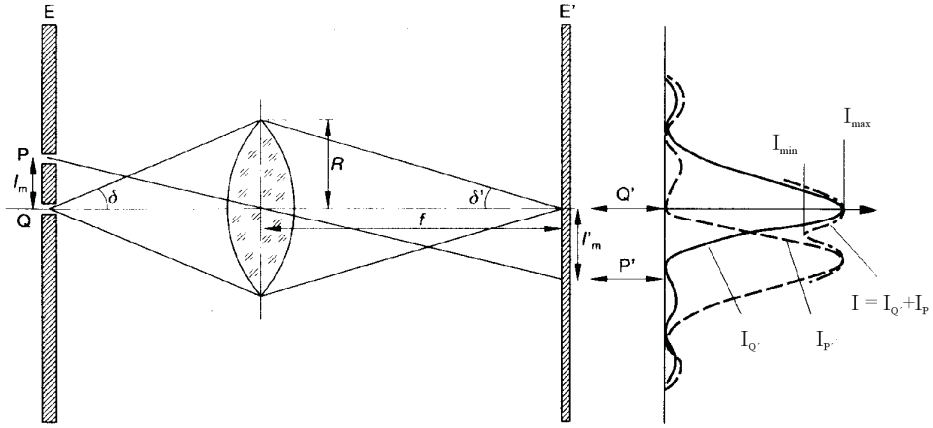


Fig. 13: *Intensity pattern of two features* P *and* Q *at projection lithography: The intensity distribution at the sample is broadened due to deflection [7]*

Resolution Enhancement

For a given tool and technology, the resolution is a given figure. There are several attempts, to improve this essential figure. When the resolution has to be improved, according to Rayleigh criterion either the wavelength λ or the technology parameter k_1 have to be decreased, or the numerical aperture *NA* has to be increased. This means either increasing the index of refrac-

tion of the medium between last optical element (by changing that medium) or physically bigger lenses. Here the problem occurs that it is difficult to produce huge lenses with the required quality; on the other hand the available materials also limit the physical size of the lenses.

To decrease the technology constant k_1, the resist can be optimized (higher contrast resist will increase resolution), or a multiple exposure can be done. E.g. the pattern of one mask layer can be segmented onto several mask layers, which can be printed consecutively. Here the interferences of light coming from structures nearby can be avoided (they are on a different mask and hence are exposed in another shot). Another method for increasing patterning performance by multiple exposures is the FLEX method (Focus-Latitude Enhancement Exposure). Here especially the DOF criterion can be overcome. In a FLEX exposure the same structure is exposed several times on the same spot but with different focus planes. This improves DOF essentially and very effective for small isolated patterns like contact holes.

Another method is altering the light path in the projection optics by introducing a pattern dependent aperture (the pupil). The method utilizes the partially coherent light. By blanking the right areas in the optical path, the interference pattern on the sample can be altered beneficially. Especially combined with off-axis exposure, this method can improve MFS and DOF.

Off-Axis illumination

In Off-Axis illumination an effective light source is created by entering an aperture between condenser and mask. The method is already known as a contrast enhancing technique for optical microscopes. With the off-axis illumination, the light beam is directed from the mask towards the edge of the projection lens, and not, as in on-axis illumination, towards the center. In normal illumination with partially coherent light, there always is a part of the light, which is off-axis, but in the context here, with off-axis illumination there is no on-axis component.

To understand the mode of operation of off-axis illumination, consider a line-and-spaces structure with pitch p. The incident light will be diffracted into a set of beams, of which only the not diffracted beam, the zero-order beam, travels in the direction of the incident light. The 1st order beam travels under the angle $|\theta_1| = \arcsin(\lambda / p)$. If p is too small, than $|\theta_{\pm 1}|$ is bigger than the acceptance angle α of the projection optics, then only the zero-order beam is projected to the sample (Fig. 14a). But this does not carry any information of the pattern, and hence the pattern cannot be transferred onto the sample. At least the zero- and the 1st order beam have to be in the range of the aperture angle. If the incident light hits the mask under an angle $\Theta_0 < \alpha$ the undiffracted beam enters the projection lens at the edge, and the 1st order beam is still collected by the lens, and therefore a pattern transfer is still possible (Fig. 14). The angle of incidence Θ_0 can be realised by bringing in the optical path an aperture between condenser and mask**.**

Although the higher resolution is an advantage of off-axis illumination, the impact on the depth off focus (DOF) is of even bigger value. In on-axis illumination, the beams of different deflection orders have to travel different ways so they are phase-shifted to each other, which results in a lack of focus. In off-axis illumination, the zero order and 1st order beam reach the projection lens at the same distance from the center, which means, that their optical path length is the same. So the relative phase difference between these beams is zero, which increases the DOF dramatically.

The off axis illumination is facilitated by an aperture (Fig. 15), which is located in front of the condenser lens. It depends on the apertures shape, which structures are improved. If there is an aperture as in Fig. 15a, only the structures perpendicular to the arrangement of the apertures will be improved. The aperture shown in Fig. 15b yields an improvement of structures,

which are adjusted to *good* angles – up/down or left/right direction. This is sufficient, because in normal cases, the features are in this *good* arrangement. The aperture in Fig. 15c even decreases this problem, but here the improvement in DOF is less.

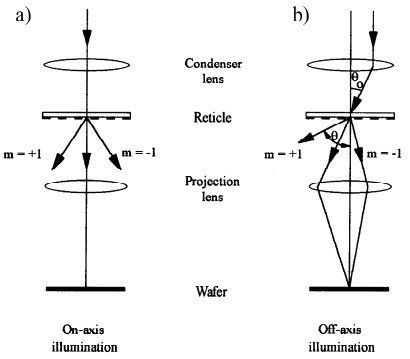


Fig. 14: *a) Optical path and deflection orders of on axis and b) off axis illumination. Note that with the same wavelength and structure size, the off axis illumination allows the 1st order beam to pass the optical system [8].*
A good description of off-axis illumination is also found in [9]

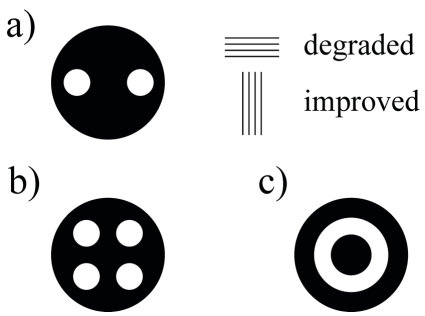


Fig. 15: *Apertures facilitating off-axis illumination. a) improvement of resolution perpendicular to the holes in the aperture, b) improvement in up-down and left-right direction, but not in diagonal direction and c) improvement in all directions [8].*

Phase Shifting Techniques

A huge improvement in resolution and/or in depth of focus can be obtained by improving the contrast by tailoring the phase differences of the wave front. The phase difference is changed by varying the optical path length of the light passing through vicinal structures, leading to constructive and destructive interference, which improve contrast (i.e. increase I_{max} or decrease I_{min}). To understand the method the approach proposed by Levenson in 1982 [10] is discussed.

Consider a lines and spaces structure with pitch 2b. Fig. 16 shows at the left hand side the amplitudes and intensities in case of a conventional mask. At the mask itself, the normalised amplitudes are of rectangular shape (either +1 or 0) and give a proper image, but the light is diffracted into the dark regions and so the amplitude distribution is broadened like shown in Fig. 16. The intensity of the light is the square of the sum of the amplitudes, so there is an intensity distribution with a significant I_{min} between the maximum intensities.

Now consider the case, when the amplitudes of the light passing through vicinal structures are out of phase by π (Fig. 16, right hand side). Again the light is diffracted into the dark areas, but now the light interferes destructively: There is a point, where the sum of the amplitudes is zero, so the intensity is zero, too. These so called *Levenson* or *alternating phase shift masks* (PSM) can improve the resolution by 40%. Unfortunately, this improvement is pattern dependent; for a single structure, there is no neighbouring structure, so there is no light to interfere with. Even if there are structures, which are not in regular arrangement, there is no defined phase shift between these structures, which could yield an improvement in the resolution of all structures.

The phase shift can be obtained by an additional transparent layer on the mask, too. If it has the refractive index n and thickness d, the phase shift is $\Phi = (n-1)2\pi d / \lambda$. So a shift of π is obtained, when the condition $d = \lambda/[2(n-1)]$ holds. On the other hand it is also possible to recess the mask material to adjust the right optical path difference. But the etch depth can be controlled by the time only, and not, as in etching away an additional layer, by the thickness of the layer itself.

To deal with the drawbacks of alternating PSM several other methods have been developed. In rim-PSM the whole mask in covered by a phase-shifter material and then with the resist. After development, the phase shifter is etched anisotropically and the masking layer is etched isotropically. By this an undercut under the phase shifter occurs at the rim of every structure. This yields a resolution improvement, too, but not as much as with alternating PSM, but therefore it is not limited to certain structures.

Another way to engineer the optical path lengths is the attenuated PSM. Here the opaque layer is replaced by a partially transparent (about 10%) layer. The light passing these semi-opaque areas is not strong enough to expose the resist, but it can interfere with the light passing the transparent areas. So an improvement of resolution can be obtained. The advantage of attenuated PSM is the easier mask processing. There is no extra layer as in alternating or rim PSM. The technology to process the semi-transparent layer is in principle the same as with the normal opaque layer.

PSM techniques were introduced since 1982, but only from 1999 they were used for industrial production.

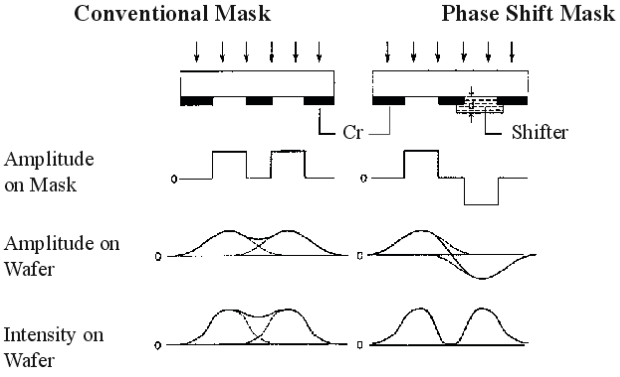


Fig. 16: *Comparison of the light amplitudes and intensities at the mask and on the wafer for a conventional and a phase shift mask. Remark that the intensity on wafer between the two features is zero for the phase shift mask [10].*

Immersion Lithography

The recent development in optical lithography using a wavelength of 193°nm or 173 nm is the altering of the numerical aperture *NA*. For *NA* is given by $sin(\alpha) \cdot n$, either α or n has to be increased. Increasing α means to increase the diameter of the optic elements. Here a physical limit is given, because it is disproportional difficult to manufacture bigger lenses with the accuracy needed. Therefore n has to be increased. For n being the index of refraction of the medium between final optical element and the sample, the medium has to be changed. This is done by flooding the gap between sample and final lens with a high index of refraction fluid, e.g. water. For n_{water} = 1.44 at λ=193 nm MFS can be improved theoretically by 44% [11]. But with increasing the *NA*, also the DOF is decreased (for water the change is about 50%), so the advantage of a higher resolution has to be paid with a lower DOF. On the other hand,

when a certain resolution is given (i.e. *NA* is constant), the DOF will be increased by a factor of *n* (cnf. eq. 3) [12].

The development of this method first dealt with the question how to avoid bubbles in the fluid, which would strongly affect the quality of the printing. Those bubbles may stem from outgassing from the resist. Under normal exposure condition, however, only low outgassing occurs [13]. Another source is the gas dissolved in the fluid itself, which may form bubbles when the fluid is heated; therefore the fluid has to be degassed. The third source of bubbles is the topography of the wafer, when the waterfront meets an air pocket (e.g. a hole on the surface of the wafer). This problem can be solved by designing the water dispensing end extraction unit of the tool to minimize such defects [14,15].

The impact of dissolving components from the resist into the fluid is low, because of the low time when the resist surface is covered with fluid [16]. It was found more problematic to minimize the effect of the fluid on the final lens. It was shown, that the lens deteriorates when exposed to the fluid. Coatings on the final lens (e.g. 200 nm of SiO_2) can act as diffusion barrier, but due to the intensive exposure to light of low wavelength, those coatings are deteriorated, too. Here still work has to be spent.

Nevertheless, by immersion lithography the optical lithography can be extended into a range, where the same results could be obtained using the 193 nm exposure wavelength as when the wavelength is changed to 157 nm. Several IC-manufacturers already have introduced or at least announced the introduction of immersion lithography into their processes.

3.2 Exposure wavelength and light sources

Progress in optical lithography was mainly achieved by decreasing the exposure wavelength λ from 436 nm to 173 nm; research is in progress to push this boundary down to a few nm. In this section the different wavelengths, the methods of obtaining that light and the implication for the process are discussed. In Table 1 the wavelengths, the sources and the name of the wavelength ranges are given.

The first light used was the light emitted from Hg-arc lamps. It provides three lines, the G-line (436 nm), the H-line (405 nm) and the I-line (365 nm). With typical k_1 and *NA* resolution of ~400 nm were achieved. Further decrease of λ to 250 nm was obtained by a mixture of Hg and Xe, improving resolution to 300 nm, but the intensity at this wavelength is low.

To solve the intensity problem, at 250 nm a new light source occurs: the excimer laser. The word stems from *exited* and *dimer* and describes a molecule, which only exists in an exited state. The gas mixture in an excimer laser is either KrF, ArF or F_2, resulting in the so called Deep UV (DUV) wavelengths of 248 nm, 193 nm and the Vacuum UV (VUV) wavelength 157 nm, respectively. The excimer molecule consists of a noble gas and a halogen atom; in ground state, they cannot react, but if one or both are in an excited state, an exotic molecule can be formed. These dimer molecules decay into the ground state of the both constituents under the emission of DUV light. The spontaneous decay time is long (i.e. nano- to microseconds), so inversion can be achieved by pumping the laser gas electrically.

Between 157 nm and 13.5 nm is a huge gap where no usable wavelength exists. This is because all materials absorb light of that wavelength, so no mask, lenses or mirrors can be made. Nevertheless, when wavelengths in the range of 13 nm are used, it is possible to set up an optical path by mirrors to do projection lithography. This range is called Extreme Ultra Violet (EUV). Shrinking λ further down to the range of 1 nm leads to x-ray lithography. Here

it is not possible anymore to do any projection lithography, because there is no material to set up an optical system.

The methods to generate this light are the same for both ranges. All of them have to meet certain requirements as being efficient enough at the desired wavelength and low debris production (or featuring a mechanism to avoid the contamination of tool and sample).

Wavelength [nm]	Source	Range
436	Hg arc lamp	G-line
405	Hg arc lamp	H-line
365	Hg arc lamp	I-line
248	Hg/Xe arc lamp; KrF excimer laser	Deep UV (DUV)
193	ArF excimer laser	DUV
157	F_2 laser	Vacuum UV (VUV)
13.5	Laser Produced Plasma; Discharge Produced Plasma	Extreme UV (EUV)
~1	x-ray tube; synchrotron	x-ray

Table 1: *Illumination wavelengths, light sources and light ranges*

Firstly an x-ray tube can be used, where a metal anode is radiated with high energy electrons, so that the characteristic x-ray radiation of the metal is emitted. The wavelength can be adjusted by the right choice of metal and electron energy. Unfortunately there are two drawbacks. On one hand, the intensity of those sources is very low leading to exposure times of several hours. This may match the requirements for research purposes, but surely not the ones of industrial production. On the other hand the anode metal is sputtered and contaminates the exposure tool and the sample. This leads to unwanted loss of intensity, which decreases the lifetime of the tool and destroys the sample. In summary x-ray tubes do not meet the requirements of industrial applications.

3.3 EUV-Lithography

EUV-Lithography is the next generation lithography (NGL). In EUV a wavelength of 13.5 nm is used. The resolution capability is determined by the Rayleigh criterion as in optical lithography. For in EUV the wavelength being so short, the resolution is enhanced even when the Numerical Aperture *NA* is lower (typically ~0.3) than in modern immersion lithography steppers (~1.2). Currently it is thought that EUV lithography can be introduced in production at the 22 nm technology node. Nevertheless the very short wavelength provokes technological challenges. Efficient light sources have to be developed. In recent years there was progress made, but still there is no light source meeting all requirements for high volume production. At such short wavelengths, there is no material transparent enough to be used for refractive optics; also mirrors cannot be made with reflectivity ~1. Mirrors can only be made from Bragg reflectors consisting of e.g. up to 40 double layers of Mo/Si, which have a reflectivity of ~70%. The mask reticle itself has to be a mirror with a patterned absorbing layer on top,

which carries the pattern information. The whole system has to be at pressures of ~10^{-3} mbar, because the absorption in air is too high.

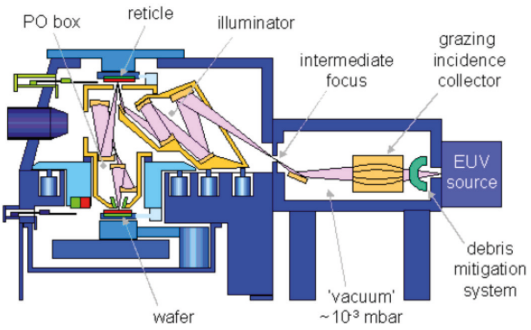


Fig. 17: *Schematic view of an EUV-scanner. The light from the source is mitigated from debris and focused to an intermediate focus. Here it enters an illumination optics, which ensures the correct illumination of the mask reticle. After the mask reticle the light is projected towards the wafer by a projection optics (PO box) [17]*

In Fig. 17 a cross section through an EUVL tool is given. The 13.5 nm radiation is emitted from point like light source. A debris mitigation system will prevent the subsequent optics from being deteriorated by the plasma of the light source. The light is focused on an intermediate focus by collector optics. In Fig. 17 a grazing incidence collector is shown, but there are other possible concepts. The subsequent illumination optics will shape the beam and assures a homogeneous illumination of the reticle. After the reticle, there is a projection optic, by which the image of the reticle is projected to the sample. Due to the low reflectivity of the mirrors, the total number is restricted. While in DUV tools there are several tenths of lenses or mirrors to minimize aberration effects, there are only ~10 mirrors in an EUV tool. Therefore EUV mirrors have aspherical surfaces to impose various aberration correction functions onto one mirror surface. The use of reflection optics in an EUV tool will restrict the numerical aperture *NA* to ~0.3, because the incident light reflected by the mirror goes to the same direction as that from which the light ray come. To avoid interference the incident and reflected light, the usable solid angle is restricted. In Fig. 18 a pre production EUV scanner is shown. The tool was developed in cooperation with the Fraunhofer Institute for Laser Technology in Aachen.

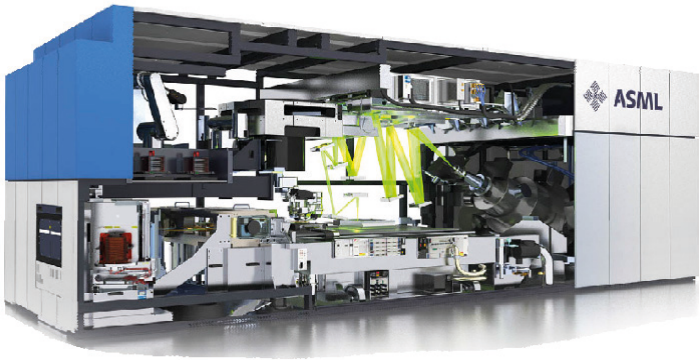


Fig. 18: *First pre-production scanner NXE:3100, using EUV source, developed in Aachen. Specified resolution: 27 nm with 60 wafers per hour.*

3.4 E-Beam Lithography

Another way to achieve sub-100nm resolution is to change the type of radiation. In the foregoing chapters only lithography methods have been discussed, which use light as illuminating radiation. It is also possible to do the illumination with charged particles as electrons. Electrons can be generated easily, either by thermionic or field effect emission, and focused to beams with a spot size of a few nanometer. This electron beam can be used to write the desired structure directly into the resist. The chemical reactions in the resists are the same as in optical resists, only the reactive species has to be attuned to the electrons (nevertheless some optical resists can be used as electron beam resists, too).

In electron beam direct write electrons extracted from the source are formed to a beam and are accelerated to a determined position on the wafer surface, where the resist has to be exposed to form the pattern. An electron beam system consists out of the electron source or electron gun, the electron-optical system (the electron column), a mechanical wafer stage and a controller system. A schematic view of an electron beam lithography tool is given in Fig. 19.

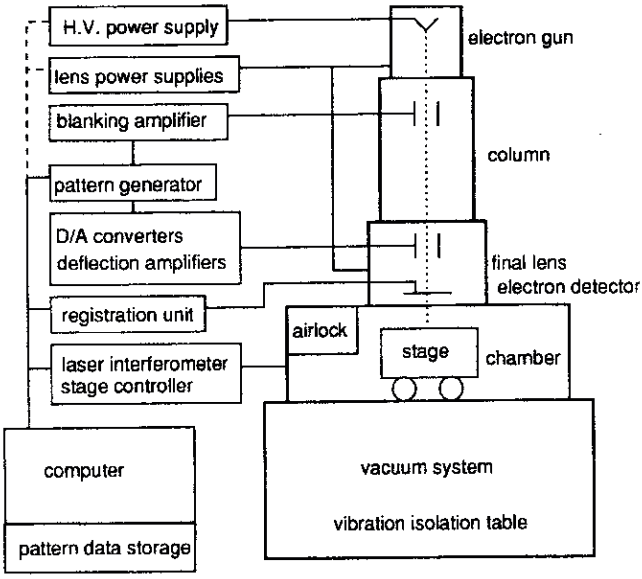


Fig. 19: *Schematic view of a Electron Beam Lithography tool [18]*

The two types of electron guns, which are commonly used, are thermionic sources on the one hand, and field emission sources on the other hand. In thermionic sources the electrons are emitted by heating the source material, as tungsten (W) or lanthanum hexaboride (LaB_6). While LaB_6 offers a higher brightness (10^5(A/cm^2)/steradian)) and a longer lifetime (~1000 hrs) as W (10^4(A/cm^2)/steradian; ~100 hrs), W has the advantage that vacuum conditions are not as high as for LaB_6. Nevertheless, LaB_6 became the standard source for thermionic E-beam sources.

In field emission sources the electrons are extracted from a sharp tip by a high electric field. Though these sources have a high brightness (10^7(A/cm^2)/steradian)), they are unstable and require an ultra high vacuum.

In the electron column the extracted electrons are formed to a beam with a definite diameter or shape. Therefore different electro-optical elements as focussing and defocusing lenses and apertures are employed. Further parts of the column are a beam blank to switch the beam on and off and a beam deflection system, with which the beam is positioned on the wafer.

Since the deflection system can only address a field of 400-800 µm (depending on spot size and tool), it is necessary to move the sample under the beam from one exposure field to the next by a mechanical wafer stage. The position of the stage is measured by an interferometer, so it is possible to adjust the beam with an accuracy of ~1 nm.

The whole system has to be under vacuum to enable the electron beam to be formed and had to be isolated form vibrations. Further requirements are low electromagnetic stray field, because this would hamper the positioning of the beam.

The pattern, which is given as a CAD file, is translated into movements of the electron beam and the wafer stage. During exposure, the distance between electron optic and sample is measured continuously and the focus is adjusted. There are three exposure schemes (Fig. 20): in the raster scan scheme, the deflection system and the wafer stage address every point of the sample, but the beam is switched on and off according to the structure. In the vector scan scheme, only the points, which have to be exposed, are addressed. Hence the vector scan scheme is less time consuming as the raster scan scheme. In both schemes the shape of the beam is point like with a Gaussian intensity distribution. In the third scheme, the shaped beam, the cross section of the beam will be shaped to different shapes. The shape itself depends on the pattern to be transferred.

Fig. 20 shows the impact of the different exposure schemes. While in raster scan mode (a) the whole wafer has to be addressed, in vector scan mode (b) on the point to be exposed are addressed, which saves time. With the shaped beam mode, the same pattern as in (a) and (b) can be exposed in only two shots, which is a tremendous saving of time.

The time needed for the illumination of a whole wafer is depended on the pattern, but because the electron beam direct write is a serial method, it is time consuming and not suitable for the industrial mass production of microelectronic circuits. Nevertheless, because the resolution is pushed to a few nanometer, it has a high impact to research activities and is used for defining the pattern on the masks, which are used for optical lithography.

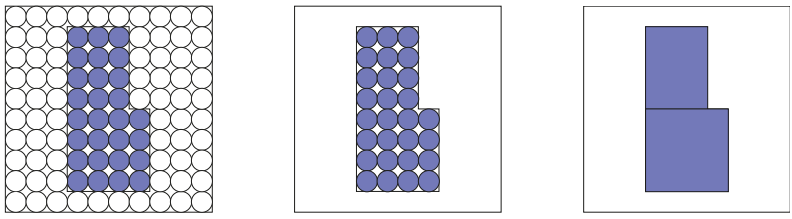


Fig. 20: *Comparison of raster scan (left), vector scan (middle) and shaped beam vector scan (right) [19]*

The resolution is not limited by the deflection of the 1–100 keV electrons, but by the beam spot size (~1 nm achievable) and by the backscattering of electrons. Fig. 21 shows Monte Carlo simulations of the electron paths, which have energy of a) 10 keV and b) 20 keV. The electrons lose their energy slowly and a significant fraction of them are backscattered to the surface,

where they expose the resist even at positions a few microns apart from the location of incidence. This so called proximity effect leads to the fact, that the effective exposure dose, with which the resist is exposed at one location, depends on the shape of the pattern in the vicinity and has to be taken into account, when the pattern are developed ("proximity correction").

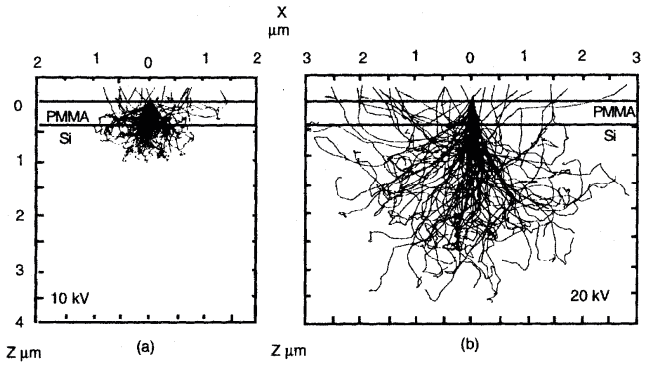


Fig. 21: *Monte Carlo simulations of the electron path in the resist and in silicon for 10kV (left hand side) and 20kV (right hand side) acceleration voltage. It is seen, that there is a significant intensity of backscattered electrons in the vicinity of the written pattern [20].*

3.5 NIL

There are several approaches for patterning structures without lithographic methods, e.g. a silicon surface can be modified by depassivation by the tunnelling current in an UHV-STM [21,22], or the surface can be modified by the movement of an AFM-tip. A certain interest has been focused on the Nano Imprint Lithography (NIL), which is described closer in this chapter.

With the NIL, a mold is processed by conventional, i.e. e-beam lithography and etching techniques and is pressed on a resist coated substrate. The structures in the mold are transfered into the resist and can be utilised after removing the mold. There are two different kinds of NIL, the hot embossing technique and a UV-based technique. A sketch of both techniques is given in Fig. 22.

Hot Embossing Technique

Here the sample is heated above the glass transition temperature of the resist, which is a thermoplastic polymer. Above that temperature the polymer behaves as a viscous liquid and can flow under pressure. The mold itself can be made of different materials, usually a silicon wafer with a thick SiO_2 layer is used. This SiO_2 layer is patterned and structured by e-beam lithography and anisotropic reactive ion etching. The aspect ratio of the features are 3:1 to 6:1, and the mold size is several cm^2. As thermoplastic polymers either PMMA or novolak resin-based resists are in use. PMMA has a small thermal expansion coefficient of $\sim 5 \times 10^{-5}\ K^{-1}$ and a small pressure shrinkage coefficient of $\sim 3.8 \times 10^{-7}\ psi^{-1}$. To ensure a proper removal of the mold, the resist is modified by release agents, which decrease the adhesion between mold and resist. Resist layers between 50 and 250 nm thickness are used. The imprint temperature and pressure depend on the resist. For PMMA the glass transition temperature is about 105°C, so the temperature, on which the sample and the mold are heated, is between 140 and 180°C. Then the mold is pressed on the sample with pressures of about 40-130 bar. Then the temperature is lowered under the glass transition temperature and the mold is removed. The features of the mold are now imprinted in the resist. The residual resist layer in these features is removed by anisotropic reactive ion etching. Afterwards, the structures can be transferred to the substrate either by di-

rect etching or by metal deposition and lift off. Structures down to a feature size of 10 nm for holes and 45 nm for mesas are imprinted with a high accuracy. [23,24]

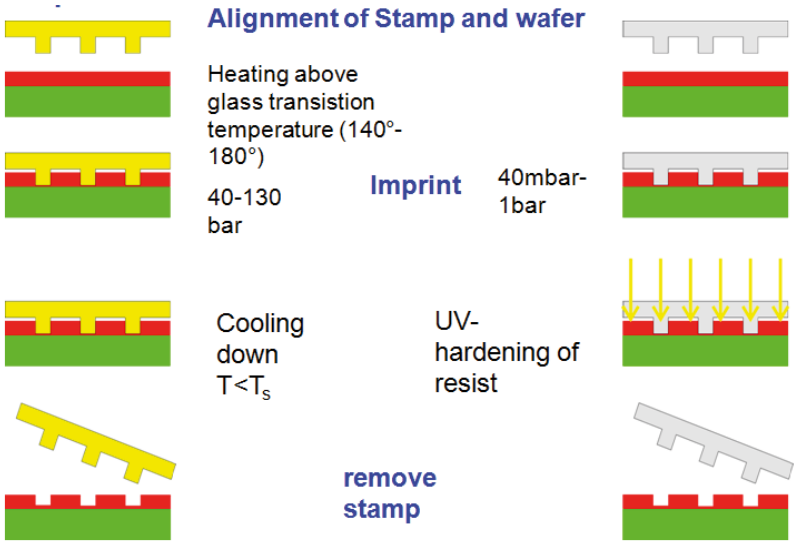


Fig. 22: *Process flows in NIL: in hot embossing (right hand side) the resist is heated above the glass transition temperature and the mold is pressed in the liquid resist. After cooling down the mold can be released leaving its form in the resist. In UV cure process the resist is fluid and the mold can be pressed with little pressure into it. The resist is hardened by UV irradiation and the mold is released.*

UV based NIL

The heating and cooling of mold and sample are time consuming. Therefore to achieve a somehow higher throughput, the curing of the resist by UV irradiation is used. Therefore the thermoplastic resist are replaced by UV-curable monomers. The mold has to be fabricated of a UV-transparent material, e.g. quartz. The features are transferred to the mold by e-beam Lithography and a Ti/PMMA resist stack. The patterned PMMA is used to transfer the features into the Ti, and the Ti is used to structure the quartz mold. The resists are acrylate- or epoxide-material systems, which can be modified concerning low viscosity, UV-curability, adhesion to the substrate and detachment from the mold. The low viscosity is essential for using low imprint pressures of 40 mbar-1 bar. After pressing the mold on the sample, the sample is irradiated by UV-radiation through the mold and a polymerisation, and hence a baking of the resist is initiated. This step lasts only about 90 seconds. After the detachment of the mold, the residual resist is removed by RIE and the further pattern transfer can be done. Again mold areas of several square centimetres can be imprinted in one run, and one imprint step last about 10 minutes. The minimum feature size reported in literature is 80 nm for dots. [25]

The NIL offers the opportunity to define dekananometer features in a rather “simple” manner, at least in comparison to advanced lithography methods described above. The field size of ~2×2cm^2 is comparable to a die, which is illuminated by a stepper. On the other hand this method is time consuming (>10min for one imprint) and till now only structures on a plain

surface are investigated, while advanced lithography is able to define structures on texturated substrates. Nevertheless, because of its technological simplicity, the NIL will be an alternative for research and small series production.

3.6 Overlay

A modern microelectronic circuit needs several mask layers, which has to be proper aligned. For that purpose, every mask layer has alignment marks: Special features on the mask, so called targets with precisely known positions and which are transferred to the sample by the subsequent etching or deposition step. The next mask layer also has alignment marks at the corresponding position. Consider an exposure tool, in which the mask is loaded in a mask mounting fixture and the wafer on a movable wafer stage.

There are two systems of alignment in use: At first, the off-axis alignment was developed. The alignment marks on the sample were observed by a separate microscope using broadband non-actinic light as illumination to prevent the resist from being exposed. The wafer alignment marks were adjusted to marks, which are etched into the microscopes objective. The mask was aligned independently to marks on the mask mounting fixture. This procedure would be enough for a single exposure, if the mask and the wafer were separately aligned properly. But what is to do, if the wafer has to be exposed by several shots as in a modern steppers? Sure, you know the structures to be transferred and you know the exposure positions on the wafer. So it is possible to move the wafer stage to every exposure position, but the movement of the stage has to be very precisely, and the long term stability of the distance between alignment position and first exposure position, the so called “base line” is difficult to achieve.

The second system is the through the lens alignment: Here, the image of the alignment mark on the sample is projected on the corresponding mark on the mask and they are compared directly. One problem occurring with this system is the alignment illumination. A He-Ne-laser is used for that, so the resist will not be exposed, but the optics is not designed for that wavelength. Therefore the lens errors for that wavelength have to be corrected by additional lenses, which are brought into the optical path.

In contact and proximity lithography, mask and wafer are aligned at every exposure. In modern projection lithography tools (e.g. a stepper), were the wafer is not exposed in one exposure, but in several shots, attention has to be paid for an accurate, quick and space efficient alignment. The wafer is moved by the wafer stage, while the mask is fixed. The position of the stage can be measured by laser interferometers very precisely. Therefore several alignment strategies have been developed.

There are three degrees of freedom: x-shift, y-shift and rotation Θ. The first steppers used a technique called “two point global exposure alignment”. In this strategy, one mark is used, to adjust x- and y-shift, and the second mark is used to adjust Θ. Afterwards, the wafer is blind-stepped through every exposure field. Here the movement of the stage has to be precisely. To avoid this problem, a “zero-level alignment mark” can be used: A mark, which is etched into the wafer before the process started. To that mark, every mask-layer is aligned.

The next approach is the “site-by-site” alignment – that means performing the alignment at every exposure position. But this is time consuming and, even a bigger drawback, there have to be alignment marks at every exposure position, which is a waste of space. Furthermore, the alignment marks on every site have to be small, so it is more difficult to detect them, resulting

in an increase in overlay error. So the “site-by-site” alignment strategy does not provide any advantage towards a global mapping strategy as the two-point-global-alignment-strategy.

In “enhanced global alignment strategy” 5 to 10 alignment marks scattered across the whole wafer are aligned and the stage position is measured several times for each marker. From the positions a least square fit is computed. Based on those data, the positions of the exposure sites are corrected. These corrections are assumed to be stable for the time, which is needed to expose the whole wafer.

With modern High Volume Manufacturing Steppers an overlay of 4.5nm ($\mu+3\sigma$) can be obtained.

The question how to find the alignment marks and how to align them proper to each other is not discussed within this work, but the reader is referred to [26] to get a survey of that theme.

While in optical systems a CCD camera will capture an image of the marker which can be handled by a pattern recognition system to search the markers automatically, the question arises, how markers are found in E-Beam-lithography. In EBL markers are square structures which show a contrast to the substrate surface for the backscattering of electrons. For marker search, the intensity of the backscattered electrons is measured. When the beam passes the edge of the marker, the slope in intensity is evaluated to extract the position of the edge. The position of the marker can be measured with sub-nm accuracy. To get overlay accuracy better than 10 nm special care has to be spent on the way how the markers are positioned on an optical mask. Avoiding exposure of the markers with a deflected beam and waiting long times after loading the specimen into the EBL system before exposure to avoid thermal drift are the key features to ensure high accuracy overlay [29]. In Fig. 23 the comparison of overlay accuracy for conventional marker definition (with deflected beam) and optimized marker definition is shown.

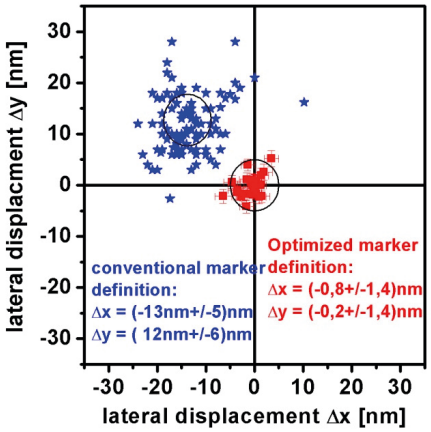


Fig. 23: *Comparison of the impact of marker definition procedure on overlay accuracy. Every symbol represents the displacement of a structure aligned to a given marker set. While in conventional marker definition the markers are exposed with deflected beams and in a random order (in respected to the markers used to align one structure), in optimized marker definition procedure, markers used to align one structure are exposed in a well defined way [29].*

4 Structure Transfer

After depositing materials and defining structures, those structures have to be transferred into the sample. This can be done by means of etching or processes like the lift off process. The

etching can be distinguished into wet chemical etching and dry etching. While in wet etching the sample is immersed into the liquid etching agent designed to remove the materials to be removed and let other parts of the sample untouched, in dry etch gases are decomposed by a plasma discharge in a vacuum chamber. The ions and free radicals will etch the material.

Etching processes are characterized by means of etch rate of the material to be etched, selectivity to the etching mask or other materials and etching profile. The etching profiles can be isotropic or anisotropic. Fig. 24a shows a schematic view of an ideal etch: the structure in the masking layer (grey) has a size of p. In an ideal etch the structure is transferred into the substrate anisotropic with a etching depth of d_e, that means the structure adopts the size p from the mask and the mask layer is not etched at all. In Fig. 24b a real anisotropic etch is shown: the etching of the substrate is anisotropic, that means that the etching is only performed in one direction (down) and not to the side. But during etch, the masking layer is thinned by d_m. The ratio d_e/d_m gives the selectivity S between material to be etched and the masking material. This etching of the masking layer also leads to erosion of the mask at the rim of the structure. During etching, the rim of the structure is pushed back by d_r. This new edge of the mask is transferred into the substrate leading to a tilted sidewall and a widening of the structure by $2d_r$. Fig. 24c shows the scheme of an isotropic etch. Here the etching takes place not in one direction only, but also to the side, leading to an under etching of the mask d_u and hence again a widening of the structure size p by $2d_u$. Depending of the application the etching method and etching agent is chosen to control selectivity S and the degree of isotropic etch (d_u/d_e).

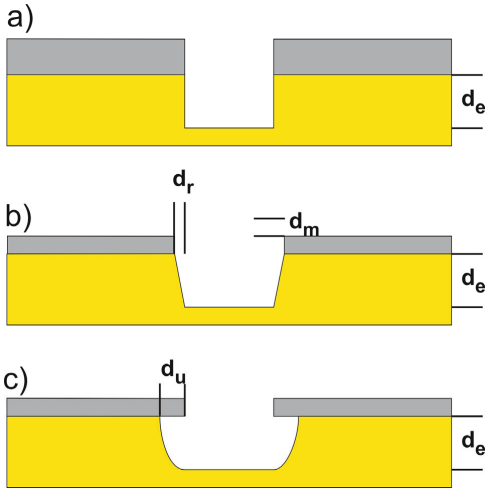


Fig. 24: *Characteristics of etching: a) ideal etch, the tech mask is not etched at all and the structure is anisotropically transferred into the sample as defined with a etch depth of d_e , b) during etch the mask erodes by the distance d_r. The structure is widened and the mask is thinned by d_m .The ratio d_e / d_m is the selectivity S of the process. c) isitropic etch: the etch front does not only proceed into the depth, but also to the side. This causes the underetch d_u of the mask.*

4.1 Wet Chemical Etching

In wet chemical etch the masked sample is immersed in a liquid etching agent. At the interface between liquid and sample a chemical reaction takes place. The process characteristics can be influenced by the temperature of the etching agent. As a prominent example for a wet chemical etching process the etching of silicon with a Si_3N_4 mask in KOH is discussed. Si reacts with strong alkaline substances (ph>12) via

$$Si + 4OH^- \rightarrow Si(OH)_4 + 4e^- \quad (9)$$

The etch rate on a crystallographic plane is determined by the material transport from and to the plane, but if there are more crystallographic planes, another effect has to be considered: the binding energy of Si is dependent on the crystal plane: While the {111} plane is only weakly etched, the {100} is etched 300 times quicker and the {110} 600 times quicker than {111} (in KOH at 85°C). While in (100) direction an etch rate of ~1.4µm/min can be obtained for Si, the etch rate of Si_3N_4 the etch rate is well below 0.1 nm/min and for SiO_2 ~1.5 nm/min. For the etch rate of a material in an etching agent is temperature dependent, the selectivity will change with temperature for the different etch rates will change differently.

Consider a Si wafer polished from both sides which is covered with a layer of 100 nm stoichiometric Si_3N_4 deposited by LPCVD (not stoichiometric SiN will show much higher etch rates). On one side, the nitride is patterned by lithography and etched with RIE. This wafer is then wet etched with KOH. If the opening in the nitride is too small, the etching will stop when the {100} plane vanishes (Fig. 25 left hand side) forming an inverted pyramid in the silicon. If the opening is well designed, it is possible to etch through the whole wafer. The {111} plane is tilted in respect to the {100} plane by 54.7°. Taking this and the wafer thickness into account, the minimum opening can be calculated (~750µm for a standard 100mm-wafer; cnf. Fig. 25 right hand side). The Si_3N_4 on the back side serves as etch stop layer.

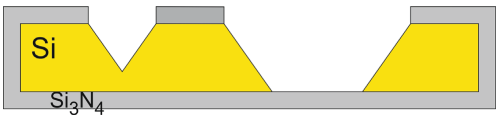


Fig. 25: *Etching of Si in KOH: using Si_3N_4 as mask, it is possible to etch through a whole wafer. If the opening in the mask is too small the etch will stop forming a inverted pyramid (left hand side). When the whole wafer is etched, the Si_3N_4 serves as etch stop forming a thin membrane.*

But what would happen when you do the same process on a (110) wafer? If you align the rim of your mask window to the (111) direction, the side wall of the etch groove will be a (111) plane and the bottom of the groove is a {110} plane. That means independent on the size of the mask window, the etching will go on with a perfect vertical side wall at the rim of the hole, until the wafer is etched through.

In general in wet chemical etching high selectivity and anisotropy can be obtained, dependent on the material system to be etched. In Fig. 26 the etch rate of $Ga_{(1-x)}Al_xAs$ in citric Acid/H_2O_2-mixture is shown. For all Al content x a minimum content of citric acid is needed to etch the material at all; above this threshold increasing citic acid content decreases the etch rate again. So a $Ga_{0.6}Al_{0.4}As$ layer in an effective etch stop for GaAs etched in 5% citric acid in H_2O_2 [30].

4.2 Dry Chemical Etch

Dry chemical etching is the most applied etching method in high volume IC fabrication, for with dry etch smaller structures can be transferred into the substrate. In dry etch the sample is placed on an electrode in a vacuum chamber. Into this evacuated chamber a gas or a gas mixture is fed through a gas supply system. This gas is used to ignite plasma by feeding an rf-

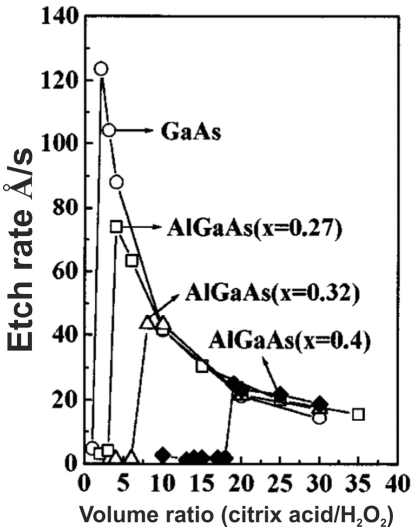


Fig. 26: *Etch rate of $Ga_{(1-x)}Al_xAs$ in citric Acid/H_2O_2-mixture: the higher the Al content x the higher the citric acid part in the etching agent has to be to etch the material. A $Ga_{0.6}Al_{0.4}As$ layer in an effective etch stop for GaAs etched in 5% citric acid H_2O_2 [30].*

power into the chamber, cracking the gas constituents into ions and free radicals. Depending on which electrode the rf-power is fed, dry etch can be distinguished into plasma etch (PE) and reactive ion etch (RIE). If the sample is placed on the grounded electrode, it is plasma etch, if the rf-power is feed into the chamber via the electrode on which the sample is placed, it is reactive ion etching. The main difference between those methods is the so called bias voltage which builds up between plasma and rf-electrode. By this bias voltage the ions from plasma are accelerated towards the rf-electrode. In case of PE this is the chamber wall and the ions do not affect the etching of the sample. Hence in PE the etching is carried by the free radicals from the plasma, which are not affected by the bias voltage. The etching process is chemically driven and has no physical component.

In RIE the ions also contribute to the etching process. While the free radicals just diffuse through the chamber to the sample, they do not have high enough kinetic energy to overcome activation energy of a chemical reaction. The ions do have a kinetic energy in the range of several hundreds of eV. So when they impinge on the sample, they can overcome activation energy and they can sputter some debris away, which would prevent the reaction. Therefore it is possible to add some gasses to the mixture, which would react with the side wall of an etched pit forming a protective layer for the reaction with the free radicals. Unfortunately this protective layer also forms at the bottom of the etch pit, which would stop the etching at all in PE, but by the physical etching by the accelerated ions, the layer at the bottom of the pit can be removed.

In a RIE process the gas mixture, the total gas flow, the sample temperature, pressure in the chamber and the rf-power are the controls with which the process is defined. As results the bias voltage, the etch rates, giving the selectivity, and the degree of anisotropy are important for the process. Increasing rf-power will increase the number of ions and free radicals, but also the bias voltage. The higher number of etching agent can increase the etch rates, but due to the higher kinetic energy of the ions also the side wall protection described above could become ineffective (which means a more isotropic etch). A higher chamber pressure would simplify the ignition of the plasma, but due to the higher ion density in the plasma there is

more scattering which diminishes the directionality of the ions, which may deteriorate the anisotropy of the process. Another degree of freedom can be introduced by a pre cracking of the gasses: the bias voltage is set by the rf-power, but if the gas is cracked before, the ion and free radical density can be controlled from the bias voltage independently. This pre cracking can be done by so called inductive coupled plasma sources (ICP). The gas is led into the reaction chamber through the top cover of the chamber, passing large coils which generate an rf-field which causes the pre cracking. The gas flows further coming in the area where the normal rf-power is applied; here the "normal" rf-power is fed into the chamber. In Fig. 27 both methods are compared.

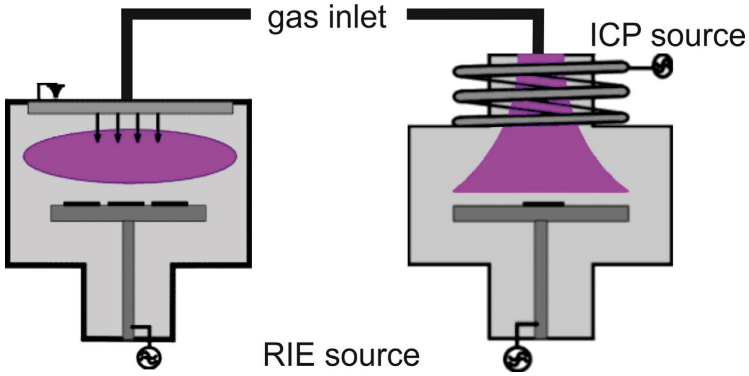


Fig. 27: *Schematic view of RIE (left hand side) and ICP-RIE (right hand side). In both cases the chamber is evacuated into the 10^{-7}mbar range. The gas mixture is fed into the chamber via the top and the gas is distributed. In RIE the gas is used to ignite a plasma by the RIE power only, while in ICP the gas is pre cracked by the ICP power. Below the ICP source the set-up of the tools are the same [35].*

The base pressure of a RIE system is in the 10^{-7}mbar range, while typical process pressures are between 5×10^{-3}mbar and 100×10^{-3}mbar, total gas flows vary between ~40sccm and ~500sccm depending on chamber size and application.

There are several effects, which may deteriorate the results of the etching: considers a silicon sample, masked with a structured SiO_2 layer as hard mask for etching. The incoming ions do have a certain x-component in velocity. On a plane surface, this will cancel out. But what happens at the rim of the mask/etched structure (cnf. Fig. 28a)? The ions with a x-component directing toward the sidewall will be reflected and will therefore reach the bottom in the vicinity of the sidewall, hence increasing the ion density locally. This increase the etch rate leading to the so called trenching effect, where a deeper trench at the rim of an etched groove occurs (Fig. 28b and Fig. 28c).

Another effect is the "aspect ratio dependent etching" (ARDE). When the ratio of mask opening (width of the structure to be etched) and the depth to be etched decreases, it comes to a depletion of etching agent at the bottom of the groove, decreasing the etch rate (Fig. 29).

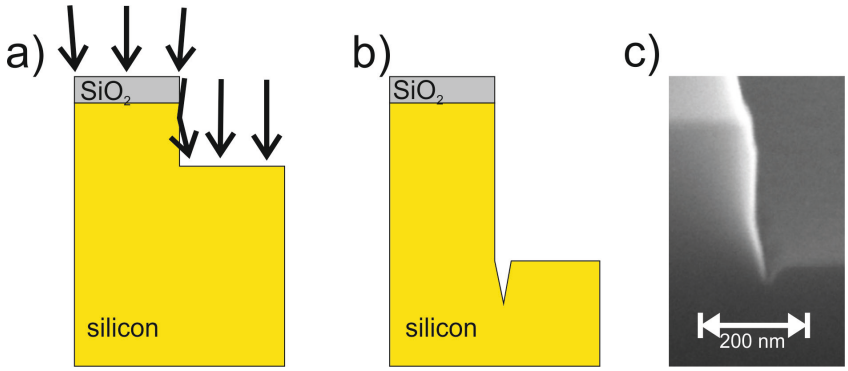


Fig. 28: *Schematic view of trenching effect during RIE: a) the ions at the sidewall of the ridge are reflected increasing the ion density and hence the etch rate at the bottom of the ridge, leading to b) a trench at the ridge. c) SEM micrograph of a trench after groove etching in Si with SiO_2 hard mask using HBr/O2 gas mixture in RIE-ICP process.*

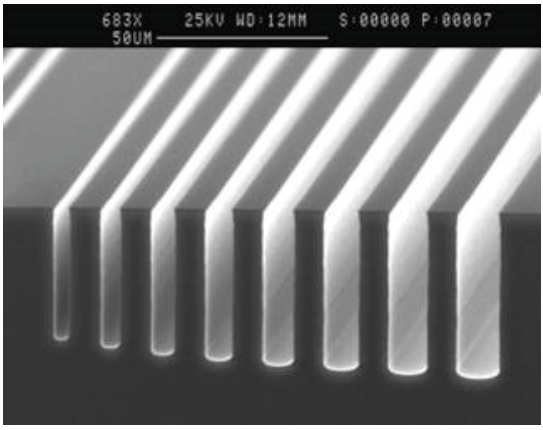


Fig. 29: *Aspect Ratio Dependent Etching (ARDE) As smaller the structure width as smaller the etch rate becomes due to depletion of etching agent [31].*

References

[1] B. E. Deal and A. S. Grove, ``General Relationship for the Thermal Oxidation of Silicon," J. Appl. Phys., vol. 36, no. 12, pp. 3770-3778, 1965.

[2] R. Singh, ``Rapid Isothermal Processing," *J. Appl. Phys.*, vol. 63, no. 8, pp. R59-R114, 1988.

[3] Richard B. Fair in "ULSI Technology" p. 149, C.Y. Chang, S.M. Sze eds. McGraw-Hill (1996) ISBN 0-07-063062-3

[4] http://www.n2bio.com/surface-modification-technology/ion-implantation.php

[5] J. Linhard, M. Scharff, H.E. Schiott „Range Concepts and Heavy Ion Ranges" Kgl. Danske Videnskab. Selbkab. Mat. Fys. Medd. 33 (1963); H. Ryssel, I. Ruge: Ionimplantation, 1. Auflage, Teubner Stuttgard, 1978; J.F. Ziegler, J.P. Biersack, J. Littmark "The Stopping and range of Ions in Solids" Pergamon Press 1985

[6] D. Klaes „Optimierung und Charakterisierung selektiv gewachsener Silizium-MOS-Feldeffekttransistoren", Berichte des Forschungszentrums Jülich; 3644; ISSN 0944-2952

[7] B. Hoppe, Mikroelektronik, Vogel, Würzburg, 1998.

[8] H.J. Levinson and A. Arnold in *Handbook of Microlithography, Micromachining and Microfabrication, Volume 1, Microlithography* (SPIE-The International Society for Optical Engineering, Bellingham, WA, 1997)

[9] B. El-Kareh, *Fundamentals of Semiconductor Processing Technology* (Kluwer Academic Publishers, 1995)

[10] M. Levenson et al., IEEE Trans Electron Dev. ED-29, 1828 (1982)

[11] A.K. Raub, A. Frauenglass, S.R.J. Brueck, W. Conley, R. Dammel, A. Romano, M. Sato, W. Hinsberg, J. Vac. Sci. Technol. **B22**, 3459 (2004).

[12] D. Gil, T.A. Brunner, C. Fonseca, N. Seong, B. Streefkerk, C. Wagner, M. Stavenga, J. Vac. Sci. Technol. **B22,** 3431 (2004).

[13] M. Rothschild, T.M. Bloomstein, R.R. Kunz, V. Liberman, M. Switkes, S.T. Palmacci, J.H.C. Sedlacek, D. Hardy, A. Grenville, J. Vac. Sci. Technol. **B22**, 2877 (2004).

[14] A. Wei, M. El-Morsi,G. Nellis, A. Abdo, R. Engelstad, J. Vac. Sci. Technol. **B22**, 3444 (2004).

[15] A. Abdo, G. Nellis, A. Wei, M. El-Morsi, R. Engelstad, S.R.J. Brueck, A. Neumann, J. Vac. Sci. Technol. **B22**, 3454 (2004).

[16] M. Sado, T. Teratani, H. Fujii, R. Iikawa, H. Iida, "Influence of water on photoresist surface in immersion lithography technology", Appl. Surf. Sci. 255 (2008) 1018-1021

[17] J. Jonkers, "High power extreme ultra-violet (EUV) light sources for future lithography", Plasma Sources Sci. Technol. 15 (2006) S8-S16

[18] M.A. McCord, M.J. Rooks in Handbook of Microlithography, Micromachining and Microfabrication, Volume 1 Microlithography (SPIE-The International Society for Optical Engineering, Bellingham, WA, 1997)

[19] L. Berger, J. Kretz, D. Beyer, A. Schwersenz, “Charged Particle Lithography“, Wiley-VCH Verlag GmbH & Co. KGaA, ISBN 9783527628155, doi: 10.1002/9783527628155.nanotech024.

[20] S.D. Berger, J.M. Gibson, R.M. Camarda, R.C. Farrow, H.A. Huggins, J.S. Kraus, „Projection electron-beam lithography: A new approach“, J. Vac. Sci. Technol. B9(6) (1991) p. 2996

[21] J.W. Lyding, T.-C. Shen, J.S. Hubacek, J.R. Tucker, G.C. Abeln, Appl. Phys. Lett. 64,2010,(1994);

[22] J.W. Lyding, K. Hess, G.C. Abeln, D.S. Thompson, J.S. Moore, M.C. Hersam, E.T. Foley, J. Lee, Z. Chen, S.-T. Hwang, H. Choi, Ph. Avouris, I.C. Kizilyalli, Appl. Surf. Sci 130, 221,(1998)]

[23] S.Y. Chou, P.R. Krauss, P.J. Renstrom: Appl. Phys. Lett. 67(21), 3114 (1995)

[24] S.Y. Chou, P.R. Krauss, P.J. Renstrom: J. Vac. Sci Technol. B 14(6), 4129 (1996)

[25] M. Bender, M. Otto, B. Hadam, B. Vratzov, B. Spangenberg, H. Kurz: Microelectronic Engineering 53 (2000) 233-236

[26] A. Moel, E.E. Moon, R.D. Frankel, H.I. Smith “Novel on-axis interferometric alignment method with sub-10 nm precision”, J Vac. Sci. Technol. B11 (1993) pp. 2191-2194

[27] J. Wensorra, M.I. Lepsa, St. Trellenkamp, J. Moers, K.M. Indlekofer, H. Lüth “Gate-controlled quantum collimation in nanocolumn resonant tunnelling transistors” nano-technology 20(2009) 465402

[28] R. Chang, C.J. Spanos “Dishing-Radius Model of Copper CMP Dishing Effects”, IEEE Transaction on Semiconductor Manufacturing 18(2), pp. 297- 303 (2005)

[29] J. Moers, St. Trellenkamp, D. Grützmacher, A. Offenhäusser, B. Rienks *“*Optimized marker definition for high overlay accuracy e-beam lithography”, Microelectronic Engineering 97(2012),pp 68–71; DOI: 10.1016/j.mee.2012.04.029

[30] J.-H. Kim, D. H. Lim, and G. M. Yang “Selective etching of AlGaAs/GaGs structures using the solutions of citric acid/H_2O_2 and deionized H_2O/buffered oxide etch” J. Vac. Sci. Technol. B, 16(2), 1998.

[31] process informations on www.oxfordplasma.de, contributed by University of Dortmund.

[32] Prof. Dr. Helmut Föll, http://www.tf.uni-kiel.de/matwis/amat/elmat_en/

[33] Stefan Peters „Rapid Thermal Processing of Crystalline Silicon Materials and Solar Cells”, Dissertation Universität Konstanz, 2004.

[34] Christian Hollauer, “Modeling of Thermal Oxidation ans Stress Effetcs”, Dissertation Technische Universität Wien, 2007. http://www.iue.tuwien.ac.at/phd/hollauer/

[35] Product information on www.oxfordplasma.de

[36] University of Glasgow, Microelectronics Process and Device Simulation Centre, http://web.eng.gla.ac.uk/groups/sim_centre/courses/semi_tech_nf.html

[37] Hong Jin Kim, Jae Kwang Choi, Myung Ki Hong, Kuntack Lee and Yongsun Ko “Contact Behavior and Chemical Mechanical Polishing (CMP) Performance of Hole-Type Polishing Pad*”,* J. Solid State Sci. Technol. 2012 volume 1, issue 4, P204-P209, doi: 10.1149/2.021204jss

[38] Prof. S. Mantl, PGI 9, Forschungszentrum Jülich, private communications

B 4 Self-Organization Techniques

Bert Voigtländer
Peter Grünberg Institut, PGI-3
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 General Principles of Self-Organization **3**

3 Self-Organization in Crystal Growth **5**
3.1 Principles of Self-Organized Crystal Growth 5
3.2 Nanostructure Formation in Heteroepitaxial Growth 11
3.3 Semiconductor Nanoislands and Nanowires 13

4 Vapour Liquid-Solid Growth (VLS) **15**

5 Bottom-Up Approaches for Resistive Switching Memories **17**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

In many areas of nature, structures form without an external guidance by self-organization. Macroscopic scale examples of such structures formed by self-organization processes range from galaxies and stars, to the formation of regular structures of clouds, sea waves, and ripple patterns in sand dunes, as the ones shown in Fig. 1. On the micro scale examples of self-organization are the formation of magnetic domains, molecular self-assembly, and crystal growth. Of course also life provides numerous examples of self-organization processes.

Fig. 1: *Ripples in sand dunes formed by self-organization processes.*

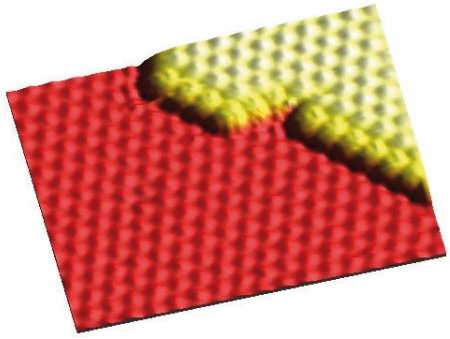


Fig. 2: *Self-organization of C_{60} clusters in a close packed hexagonal lattice due to steric interactions. Scanning tunneling microscopy image (15 nm x 10 nm).*

An essential feature of self-organization processes is the formation of ordered structures without external guidance. Typically, one can describe any self-organization process by driving forces which originate from

- thermodynamics
- kinetics
- properties of building blocks

or a combination of these. Throughout the text, we will discuss these ingredients. In many cases self-organized structures are thermodynamically stable. In closed systems a driving force for self-organization is the minimization of the (free) energy.

There are numerous cases of dynamic self-organization in open systems far from thermodynamic equilibrium. Some examples are convection cells in a gravity field found in clouds or in a fluid heated from below, in which regular Bénard convection cells form. In these cases kinetics and transport phenomena control the self-organization process. Often, this can be described by an interplay of kinetics and thermodynamics such as in the ripple pattern of sand dunes, mentioned above, which are formed by the kinetic energy of the wind, and the effect of gravity and friction.

Other key ingredients for self-organization are the building blocks which form the self-organized structures. These can range from stars and galaxies as building blocks of the universe down to the very atoms which act as building blocks for instance in self-organization processes such as the crystal growth forming regular crystal lattices. An example where steric effects (each building block requires

a certain amount of space) lead to the formation of a regular closed packed hexagonal two-dimensional lattice of C_{60} clusters is shown in Fig. 2.

As we have seen the term self-organization spans a very broad range of processes occurring in nature. In a narrower sense which is often used, the term self-organization is identified with non-equilibrium processes, whereas the term self-assembly is often used for processes proceeding under equilibrium.

Within the context of nanoelectronics there are several approaches to use principles of self-organization in order to fabricate desired functional nanostructures bottom up out of single atoms or molecules in parallel and without the size limitations set by current lithography techniques [1]. An advantage of lithographic top-down methods is the ability to fabricate a large variety of defined structures. The bottom-up methods offer the opportunity to fabricate structures in the single-digit nanometre range enabling the formation of billions of nanostructures with control over size, shape and composition in a fast and parallel fashion, but it is still a great challenge to arrange atoms or molecules into desired structures.

In this chapter we concentrate on the formation of self-organized nanostructures using different growth techniques. First we discuss general principles of self-organization and their application to crystal growth. Subsequently, we discuss the vapour liquid-solid growth method (VLS) for the formation of nanowires and bottom-up approaches for resistive switching memories

2 General Principles of Self-Organization

As has been seen above, the spontaneous formation of ordered structures is quite common and can be observed in many physical, chemical and biological systems. There are some general principles which govern the formation of ordered structures independent of the nature of the building blocks. The direction of spontaneous changes in any thermodynamic system is determined by the second law of thermodynamics which states that spontaneous processes must lead to an increase in the entropy of the universe:

$$dS_{\text{universe}} \geq 0. \tag{1}$$

The entropy of the universe, S_{universe}, might be divided into two contributions: entropy of the system under consideration, S, and entropy of the environment, S_{env}. When a system transforms from a disordered to an ordered state, its entropy usually decreases. Therefore, in a spontaneous ordering process the system must supply some amount of heat to surroundings, thereby increasing entropy of the environment S_{env} and overcompensating the entropy decrease due to ordering.

The source of the heat is the internal energy of the system, E_{int}. By decreasing its internal energy the system may produce heat and perform work. This statement is expressed by the first law of thermodynamics:

$$dE_{\text{int}} = \delta E_Q + \delta E_{\text{mech}}. \tag{2}$$

Here δE_{Q} is the amount of heat supplied to the system, δE_{mech} is the mechanical work.

In many practical situations some of the thermodynamic parameters of the system are fixed. For instance, self-assembled molecular films and self-organized epitaxial nanostructures, which are the topic of this chapter, are usually fabricated at constant temperature and volume.

When the volume is fixed, the system does not perform mechanical work and the whole change of the internal energy of a closed system, $dE_{\text{int}} < 0$, is completely transformed into heat. In this case the entropy of the environment is increased by $dS_{\text{env}} = - dE_{\text{int}}/T$ and the total change in entropy of the universe will be $dS_{\text{universe}} = d(S_{\text{env}} + S) = - d(E_{\text{int}} - TS)/T$. The quantity in the brackets, $H = E_{\text{int}} - TS$, is called the Helmholtz free energy. Since, by the second law of thermodynamics, $dS_{\text{universe}} \geq 0$, the Helmholtz free energy H of a system at constant temperature and volume must decrease in spontaneous processes, so that the system evolves to a state of minimal H.

Thus, a minimum of the Helmholtz free energy determines the equilibrium state at constant T and V. However, generally, there might be several minima of the free energy of the system. In these cases the system might be trapped in metastable states corresponding to a local minimum of the free energy. The ability of the system to reach the global minimum is determined by the height of the free energy barriers separating local and global minima and by kinetic rates of the processes changing the system state (Fig. 3).

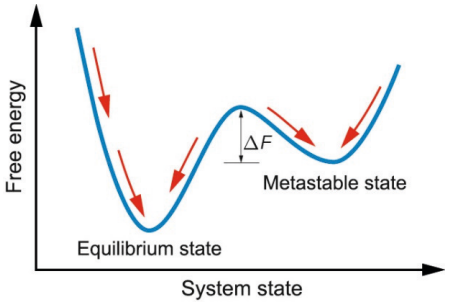


Fig. 3: *Free energy of a system for different states of the system. The system evolves towards lower free energy. The equilibrium state is the state of lowest free energy. However, the system might also be trapped in a metastable state.*

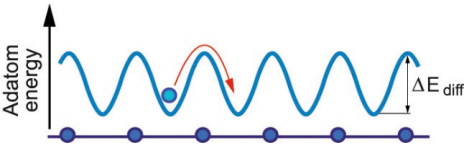


Fig. 4: *Potential seen by an atom adsorbed on a crystalline surface.*

The rates of spontaneous processes are treated by irreversible thermodynamics, which introduces thermodynamic fluxes as the rates at which processes proceed and postulates a linear relation between the thermodynamic fluxes and the thermodynamic forces that drive the system to an equilibrium state. One of the most important thermodynamic fluxes in the context of the present chapter is the diffusion flux, which is determined as the amount of matter flowing through a unit of area per unit time. The driving force for the matter transport in the system is the gradient of concentration of the diffusing particles. Neglecting the cross-effects, i.e. the influence of other thermodynamic forces on the matter flow, the diffusion flux can be written as

$$\boldsymbol{J}_{\text{diff}} = -D\nabla c. \tag{3}$$

Here, c is the concentration of the diffusing particles, and D is the diffusion coefficient. Equation (3) is widely known as the first Fick's law of diffusion.

Macroscopic changes in a thermodynamic system stem from multiple atomic-scale kinetic processes. Kinetic processes involved into the formation of self-organized nanostructures on surfaces typically include adsorption of atoms and molecules on the surface, desorption from the surface, surface chemical reactions, diffusion of adsorbed particles on the surface, formation of atomic and molecular clusters, etc. An important feature of these processes is that most of them are thermally activated. For instance, when an atom or molecule arrives at a crystal surface, it is trapped by a periodic potential imposed by the crystal. Minima of that potential form a regular network of adsorption sites on the surface. In order to move from one adsorption site to another, an adsorbed particle has to overcome an energy barrier, as shown in Fig. 4. The probability that an adsorbed atom or molecule will attain the necessary energy by thermal fluctuations is given by the Boltzmann factor $\exp(-\Delta E_{\text{diff}}/k_{\text{B}}T)$, so that the frequency of diffusion jumps of an adsorbed particle is given by $\nu=\nu_0 \exp(-\Delta E_{\text{diff}}/k_{\text{B}}T)$. In a similar way kinetic rates of other thermally activated elementary atomic-scale processes could be defined.

Since the thermally activated surface processes slow down drastically at low temperatures, an equilibrium state of the system might not be achieved on the time scale of the experiment. In some cases kinetic limitations might be intentionally used for fabrication of non-equilibrium nanostructures. Some examples of such nanostructures will be given in Sec .3. In general, the nanostructure formation represents a subtle interplay between thermodynamics and kinetics, so both have to be carefully taken into account.

The inherent properties of the building blocks forming self-organized structures are the third essential parameter of particular importance for the self-organization process. The size and the shape of individual composites as well as the bonding forces between them determine the formed architecture and the structural design. In order to tailor self-organized assemblies the following properties are important:

- The scale and the shape of the building-blocks responsible for size and packing-density of the assembly.
- Attractive and repulsive interactions between building blocks and between building blocks and environment (air, liquid, solid surface). These interactions determine the geometry and the distances at which building blocks come to equilibrium in a self-assembled system.
- The homogeneity of the constituting components should also be considered. Components are only really monodisperse, if they are atoms or molecules. In all other cases slight deviations from building block to building block have to be taken into account, which reduce the achievable perfection of the assembly and enhance the population of defects.

3 Self-Organization in Crystal Growth

One approach for the fabrication of nanostructures is epitaxial growth, i.e. a fixed epitaxial relation exists between the crystal structures of substrate and the grown epitaxial layer. Depending on the particular growth system, the growth may take place either under highly non-equilibrium (kinetic) or under (near) equilibrium conditions and nanostructures as islands can be formed. In the non-equilibrium growth the sizes of the nanostructures can be tuned down to the single-digit nanometre range by choosing appropriate growth conditions. However, the size uniformity is the greatest challenge. If the growth takes place under (near) equilibrium conditions, then the formation of nanostructures is governed by energy minimization and good size uniformity is expected. As an example of nanostructures grown by epitaxy, the formation of islands and wires will be presented. Subsequently, the growth of nanostructures on template substrates structured by step arrays or underlying dislocation networks will be considered.

3.1 Principles of Self-Organized Crystal Growth

Growth Techniques

The main methods used for epitaxial growth are chemical vapour deposition (CVD) [2] and molecular beam epitaxy (MBE) [3], [4]. In CVD growth gases containing compounds of the elements to be deposited are introduced into the growth chamber. When the gas molecules hit the substrate surface, they decompose (partially) and the chemical elements forming the growing film stick to the substrate. Different chemical reactions taking place at the surface or even in the gas phase lead to a quite complex nature of the fundamental processes of epitaxial growth in CVD. Molecular beam epitaxy is conceptually simpler. Here the elements to be deposited are heated in evaporators until they evaporate. The beam of the atoms hit the surface and the atoms diffuse over the surface and finally bind at surface lattice sites (Fig. 5).

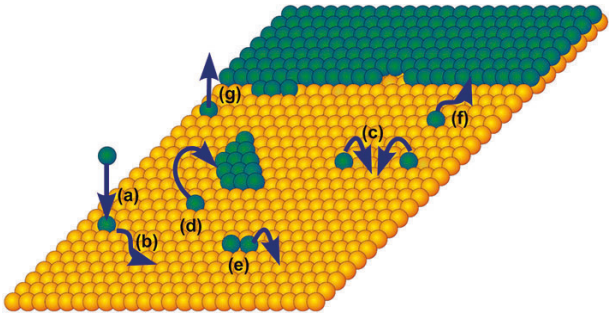


Fig. 5: *Fundamental surface processes occurring during epitaxial growth.*

In spite of the fact that the MBE growth is in principle much easier than the CVD growth, there are still a lot of different fundamental processes that occur during epitaxial growth by MBE [5]. Part of them are shown in Fig. 5. Atoms from the molecular beam arrive at the crystalline surface (a) and diffuse over the surface by performing thermally activated random jumps from one adsorption site to another (b). When two atoms (or sometimes also more than two atoms) meet, they form a cluster on the surface (c). The cluster can either grow to larger sizes by capturing further migrating adatoms (d) or dissociate by breaking bonds between the atoms (e). The cluster for which the probabilities to grow or decay are equal is called the critical nucleus [6]. Clusters which are larger than the critical nucleus are called stable islands. Clusters smaller than the critical nucleus are called sub-critical nuclei. For instance, if the critical cluster size is $i^* = 3$ atoms, dimers and trimers may dissociate and the clusters of four atoms represent the smallest stable 2D islands. Other important growth processes are the attachment of adatoms at pre-existing steps (f) and desorption of adatoms from the surface (g).

Kinetics of 2D Island Nucleation

Two-dimensional (2D) islands, i.e. islands of one atomic layer height, represent the simplest example of the self-organized nanostructures. The density of 2D islands on the surface and the island size distribution are governed by energetic barriers such as the barrier for the surface diffusion of adatoms, and, additionally, by outer conditions such as the substrate temperature T and deposition flux F. The strong temperature dependence of the island density is illustrated in Fig. 6(a) and (b) for the growth of Si on Si(111).

At low deposition temperatures typical for MBE growth of semiconductor materials, no bond breaking between atoms occurs on the time scale of the experiment. In this case the critical nucleus size is one atom and the cluster of two atoms (dimer) represents the smallest stable 2D island. The frequency with which two adatoms meet each other (the nucleation rate) depends on the adatom density n as n^2, and also on the rate of adatom surface diffusion which is characterized by the surface diffusion coefficient D:

$$\frac{dN}{dt} = Dn^2. \tag{4}$$

The total island density N can be found from (4) by integration. However, since the island nucleation rate depends strongly on the adatom density n, Eq. (4) has to be coupled with an appropriate equation for n.

One possible approach is to use the mean field approximation assuming that the growing islands are surrounded by a spatially uniform adatom field. To illustrate this approach we will neglect desorption of adatoms from the surface, thus considering the so called complete condensation limit. The mean-field adatom density n is determined then by the following rate equation

$$\frac{dn}{dt} = F - 2Dn^2 - DnN, \tag{5}$$

where the first term, F, is the deposition flux of atoms adsorbing on the surface, and the second and third terms are the loses of adatoms due to the island nucleation and incorporation of adatoms to the existing islands, respectively. The factor of two in the second term accounts for the fact that two adatoms are lost per nucleation event.

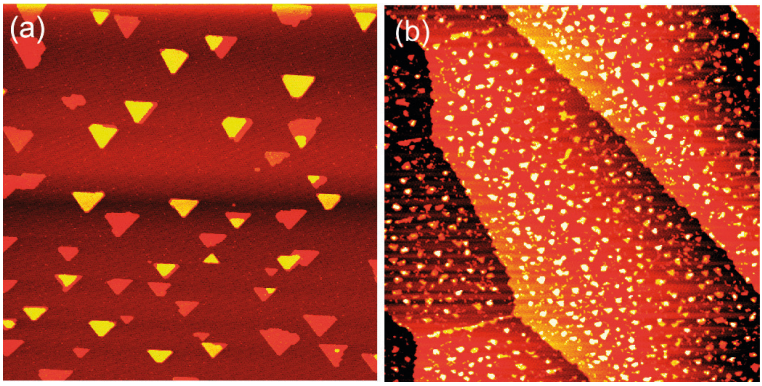


Fig. 6: *Scanning tunneling microscope images after the growth of 0.2 atomic layers of silicon on a Si(111) surface. The islands have triangular shape due to the symmetry of the substrate and have a height of one atomic layer (orange) or two atomic layers (yellow). The island density depends on the temperature, as can be seen by comparison of growth at high temperatures of 770 K (a) to growth at a lower temperature 610 K (b). Both images have a size of 350 nm.*

It follows from Eq. (5) that the temporal dynamics of the nucleation process may be divided into two stages. In the transient nucleation regime, which takes place at the very beginning of the deposition process, the loss terms in (5) are negligible, so that adatom density increases with time as Ft. This leads, according to (4), to a rapid increase of the island nucleation rate and the density of stable islands on the surface. When the density of stable islands becomes appreciable, the adatom density n drops down and the system enters the steady-state growth regime, where nucleation events are rare and most of the depositing adatoms join the existing islands. In this regime the nucleation term $2Dn^2$ is negligibly small and the adatom density remains nearly constant ($dn/dt \approx 0$) because the deposition flux is completely balanced by the incorporation of adatoms into the islands ($F \approx DnN$). The corresponding steady state adatom density can be written as:

$$n = \frac{F}{DN}. \tag{6}$$

Putting (6) into (4) and switching to a dimensionless time variable $\theta = Ft$ (total surface coverage) one writes

$$\frac{dN}{d\theta} = \frac{F}{DN^2}. \tag{7}$$

Integration of (7) yields a very useful scaling relation which links the density of 2D islands with the experimentally controlled parameters, namely the deposition flux F and temperature dependent surface diffusion constant D:

$$N \sim \left(\frac{F}{D}\right)^{1/3}. \tag{8}$$

The temperature dependence of the surface diffusion coefficient $D = D_{0_0} \exp(-\Delta W_{diff}/k_B T)$ is determined by the activation energy for adatom surface diffusion ΔW_{diff}, which is the main material-specific kinetic parameter controlling the nucleation process. The scaling relation (8) represents a special case of the well known Venables formulae [6] for the critical nucleus size $i^* = 1$. If the critical nucleus size is different from one, a generalized form of Eq. (8) can be obtained as

$$N \sim F^{\chi} \exp\left(\frac{W}{k_B T}\right), \tag{9}$$

where the energy parameter W and the scaling exponent $\chi = i^*/(i^*+2)$ are functions of the size of the critical nucleus [6]. If adatom incorporation into the 2D islands is hindered by kinetic constraints the scaling relation (9) still holds, however the scaling exponent may take different values depending on the actual mechanism of the island nucleation [7]-[9].

The importance of the scaling relations (8) and (9) is that they show how the island density can be controlled by adjusting the growth temperature and deposition flux. An example for this scaling is shown in Fig. 7 where the density of Si islands on Si(111) is plotted as function of temperature and deposition flux. A fit of the experimental data with (6) and $\chi = i^*/(i^*+2)$ yields $i^* \approx 6$, i.e. the critical nucleus for this particular system in the given range of the deposition conditions is 6 atoms.

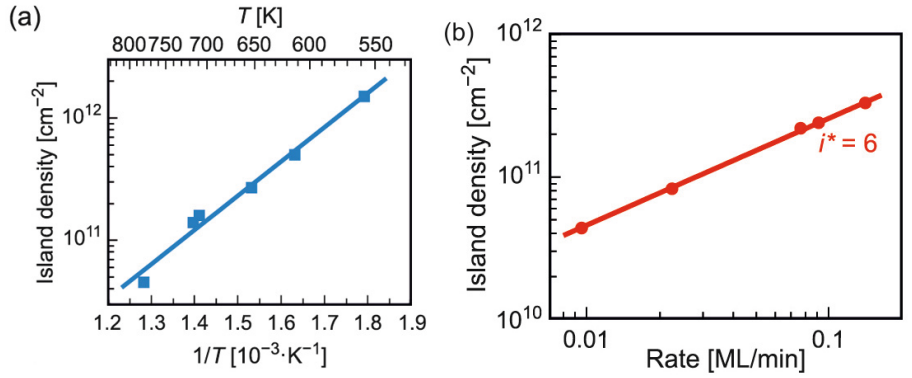


Fig. 7: *The island density of Si islands on Si(111) shows the scaling behaviour as derived in Eq. (9). (a) scaling behaviour as function of temperature (b) scaling behaviour as function of the deposition flux. The line in (b) corresponds to i* = 6. This shows that the island density can be controlled by the kinetic parameters temperature and deposition flux.*

Although the nucleation of the islands is a random process, the distribution of the island sizes is often centred around a mean value (Fig. 8). This arises due to a saturation of the island nucleation as will be explained in the following. In the early stage of growth (nucleation regime) islands nucleate randomly on the surface and the distance between the islands decreases. If the distance between the islands becomes about twice the mean distance which an adatom travels before a nucleation event happens, then the incorporation of adatoms in existing islands becomes a more probable event than the nucleation of new islands and around each island a "capture zone" forms. All adatoms deposited in this capture zone attach to the corresponding island and nucleation of new islands ceases. Without this effect the distribution of island sizes would be even broader. The island size distributions for two different temperatures are shown in Fig. 8. It is seen, that the peak in the island size distribution scales towards larger sizes with higher temperatures. This is directly related to the fact that the capture zones of 2D islands

become larger at higher temperature, because the diffusion flux of adatoms increases. In some cases also the surface reconstruction [5] or chemical passivation of the surface by a surfactant layer [10] can modify the island size distribution. Generally, in MBE growth the density of two-dimensional islands can be very accurately controlled by the kinetic parameters, temperature and deposition flux. Additionally, average island size can be controlled by the deposited amount. However, the island size distribution is quite broad due to the stochastic nature of the nucleation of the islands.

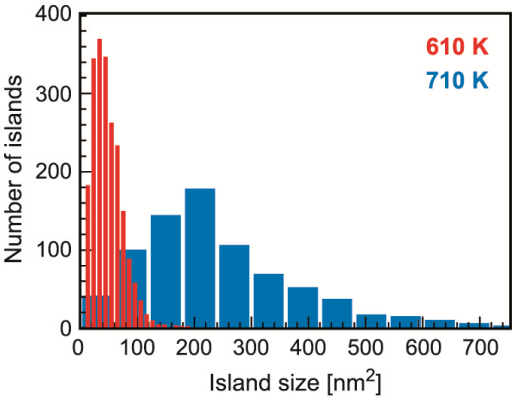


Fig. 8: *Island size distribution for two-dimensional Si islands on Si(111). The width of the distribution is of the order of the average size of the islands. Two distributions for two different temperatures are displayed. The narrow bins (peak at small island sizes) correspond to deposition at 610 K. The distribution with the wide bins (peak at larger island sizes) corresponds to deposition at 710 K.*

Thermodynamically Stable Nanostructures

If nanosized islands were thermodynamically stable their size distribution would be narrow. A thermodynamically stable island size means that the energy (per atom) has a minimum for this stable size. For configurations with larger or smaller islands the energy (per atom) would be higher. Therefore, one has just to approach thermodynamic equilibrium to obtain a very narrow island size distribution. One way to achieve thermodynamic equilibrium is to heat a sample with different island sizes present and wait until equilibrium has established. The equilibrium configuration will be established by material transport between the islands. Atoms will detach from islands with higher energy and attach to islands with a lower energy (per atom). However, as we will show below, in the simplest case (considering only a surface or edge energy term) the thermodynamically stable island size is infinitely large. This behaviour is not of any use for the formation of nanostructures with a narrow size distribution and corresponds to the ripening of larger islands in expense of smaller ones. Only if additional terms in the energy are important such as strain energy for example, the energy per particle can have a minimum for a finite particle size. In this case a narrow size distribution can be expected under equilibrium conditions.

To describe material transport in a system with a variable number of atoms the chemical potential is used; this is the change of the energy E (of an island) when the number of particles (N) changes $\mu = \mathrm{d}E/\mathrm{d}N$. During the equilibration process atoms detach from islands where the chemical potential is highest and attach to islands with a lower chemical potential. This lowers the total energy of the system. Therefore, the material transport between different islands is governed by the chemical potential. A simple example is the chemical potential of square 2D islands of dimension L (Fig. 10a). Neglecting the constant binding energy of the substrate, the energy difference between different sized islands comes from the edge energy (β is the edge energy per length). The energy of an island is $E = E_{\mathrm{edge}} = 4\,L\beta$. The number of atoms in an island (N) depends on the dimension L as $N = L^2 / \omega$ with ω being the area per atom. The chemical potential is then

$$\mu = \frac{dE}{dN} = \frac{2\omega\beta}{L} \sim \frac{1}{L}. \qquad (10)$$

Since μ is decreasing for larger islands infinite size islands have the lowest chemical potential (Fig. 10b). This means that the stable island is infinitely large. In this case the equilibration does not result in a stable finite island size. This equilibration by material transport between islands is also called coarsening because it results in the shrinkage of small islands and a growth (coarsening) of large islands (Ostwald ripening).

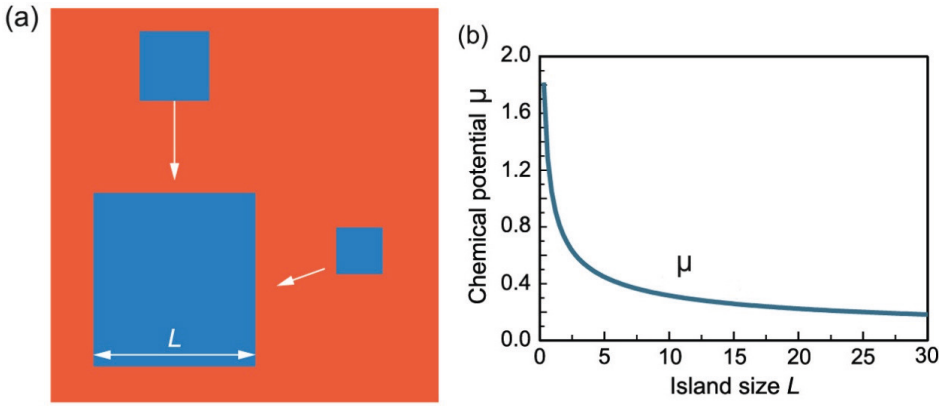


Fig. 9: *(a) Coarsening of a large island at the expense of small ones. (b) Chemical potential of an island.*

An infinitely large stable island size is the result of homoepitaxial growth, taking only into account the edge energy. The situation becomes different when also elastic stress is taken into account, as it occurs in heteroepitaxy where two different materials grow onto each other. Here, stress is induced by the different lattice constants of the substrate material and the material of the islands. Atoms in a flat continuous heteroepitaxial layer are forced to maintain the lattice constant of the substrate. However, when a heteroepitaxial island is formed, atoms at its edges are less confined and may take more relaxed positions. Therefore, the island formation in heteroepitaxial growth leads to the strain relaxation. An energy gain due to the strain relaxation by a square 2D island can be calculated using the elastic theory as $E_{\text{strain}} = 2LC \ln L$, where C is a constant [5]. Substituting this strain relaxation energy from the step edge energy results in a total energy of a strained heteroepitaxial island as

$$E = E_{\text{edge}} - E_{\text{strain}} = 2L\left[2\beta - C \ln L\right]. \tag{11}$$

This results in the following chemical potential

$$\mu = \omega\left[\frac{2\beta - C}{L} - \frac{C}{L}\ln L\right], \tag{12}$$

which is illustrated in Fig. 10. In this case the chemical potential has a minimum at the size $L_{\text{min}} = \exp(2\beta/C)$. This would mean that during coarsening the islands would approach this size. Larger islands would dissolve and smaller islands grow until all islands have the size L_{min}, i.e. lowest chemical potential. This would result in a very narrow size distribution. Unfortunately, step energies are not known very well, so that it is not possible to predict a reliable number for the equilibrium island size. An experimental realization of thermodynamically stable islands has not yet been confirmed apart from surface reconstruction domains with a relatively large unit cell.

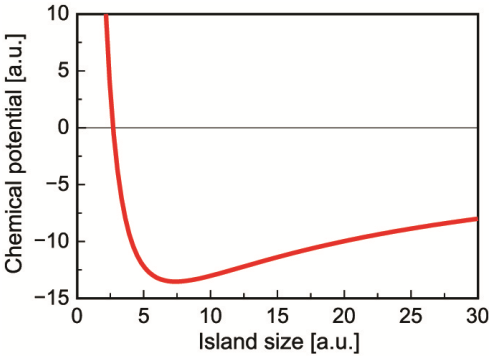


Fig. 10: *: Chemical potential of an island with an energy component due to elastic strain included.*

If we compare the formation of nanostructures in equilibrium to the formation of nanostructures by growth kinetics the following advantages and disadvantages occur. Nanostructures grown under equilibrium conditions have (under specific conditions) potentially the advantage of a narrow size distribution around the optimum size. A disadvantage is that the size is determined by the material parameters (strain energy and step edge energy for instance) and cannot be tuned freely. The size and density of nanostructures formed under kinetic conditions can be tuned easily by variation of the growth parameters such as growth rate and temperature. On the other hand the size uniformity of the islands grown under kinetic conditions is relatively poor.

3.2 Nanostructure Formation in Heteroepitaxial Growth

Semiconductor nanostructures can be fabricated by self-organization using heteroepitaxial growth which is the growth of a material B on a substrate of different material A. Since in the case of Ge/Si, which we use in the following as an example, the surface energy of the epitaxial film (Ge) is lower than the surface energy of the substrate, there is an energetic driving force for the epitaxial layer (Ge) to spread out over the Si substrate. This two-dimensional layer is also called wetting layer, because Ge "wets" Si. In addition to the surface energy, in heteroepitaxial growth, often the lattice constants of the two materials are different. The lattice mismatch for the two most commonly used material systems

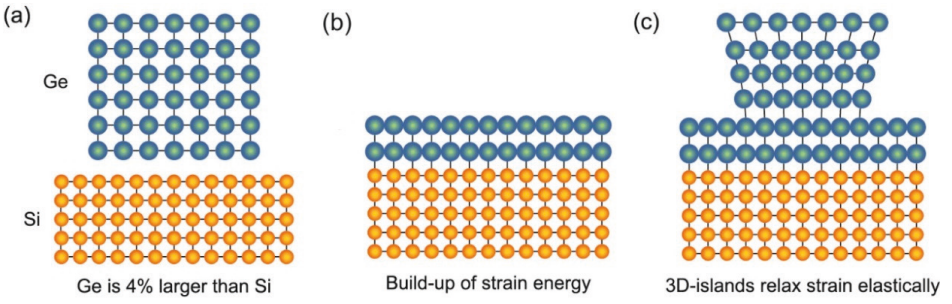


Fig. 11: *(a) Schematic representation of Si and Ge crystals with different lattice constants, (b) build up of elastic strain energy during 2D growth with Ge confined to the Si lattice constant and (c) elastic relaxation by formation of 3D islands (Stranski–Krastanov growth). In the upper part of the 3D island the lattice constant relaxes towards the Ge bulk constant. The usual form of the 3D islands is a pyramid and not like the one shown in this schematic sketch.*

Si/Ge and GaAs/InAs is 4.2 % and 7 %, respectively (schematically shown in Fig. 11a). This lattice mismatch leads to a build up of elastic stress in the initial two-dimensional growth in heteroepitaxy. In the case of Ge heteroepitaxy on Si the Ge is confined to the smaller lattice constant of the Si substrate i.e. the Ge is strained to the Si lattice constant (Figure 12b). One way to relax this stress is the formation of 3D Ge islands. In the 3D islands only the bottom of the islands is confined to the substrate lattice constant. In the upper part of the 3D island the lattice constant can relax to the Ge bulk lattice constant and reduce the stress energy this way (Figure 12c). The growth mode, characterized by the formation of a 2D wetting layer and the subsequent growth of (partially relaxed) 3D islands, is called Stranski–Krastanov growth mode. Examples are shown in section 3.3.

The driving force for the formation of self-organized 3D nanoislands in heteroepitaxial growth is the build up of elastic strain energy in the stressed 2D layer. As a reaction to this, a partial stress relaxation by the formation of 3D islands can lower the energy of the system. The process of island formation close to equilibrium is a trade-off between elastic relaxation by formation of 3D islands which lowers the energy of the system and an increase of the surface area which increases the energy.

In a simple model, where the islands are just cubes with the length x, the additional surface energy for a film in an island morphology (compared to a strained film) is proportional to the island length squared (x^2). The gained elastic relaxation energy compared to that of a flat strained film is in the simplest assumption proportional to the volume of the island (x^3). For the same total volume in the film, the energy difference between the three-dimensional island morphology and the flat morphology is

$$E = E_{\text{surf}} - E_{\text{relax}} = C_1 \gamma x^2 - C_2 \varepsilon^2 x^3, \tag{13}$$

with γ being the surface energy, ε being the lattice mismatch, and C_1 and C_2 constants. The contributions of E_{surf}, E_{relax} and the total energy difference between the three-dimensional island morphology and a flat film are shown in Fig. 12 as a function of the island size x. For small sizes of the three-dimensional islands the three-dimensional island morphology is unfavourable until to the point where the absolute value of the gained elastic relaxation energy ($\sim x^3$) becomes larger than the cost of surface energy ($\sim x^2$). For islands larger than a critical island size x_{crit}, the formation of three-dimensional islands is energetically preferred over the two-dimensional film morphology. While this simple model shows the basic driving forces for the two-dimensional to three-dimensional transition, it contains several simplifications. For instance, in this simple model the island morphology is assumed as being cubes, which does not correspond to the experimentally observed island shapes. Further, the simple model contains only energetic considerations of two final states. Kinetic effects, like the required material transport necessary during the two-dimensional - three-dimensional transition are not considered.

Apart from the formation of 3D islands there is another process which can partially relax the stress of a strained 2D layer: the introduction of misfit dislocations. This corresponds to the removal of one lattice plane in a compressively strained layer. If a lattice plane is removed in regular distances in the strained layer a misfit dislocation network forms. Depending on the growth parameters temperature

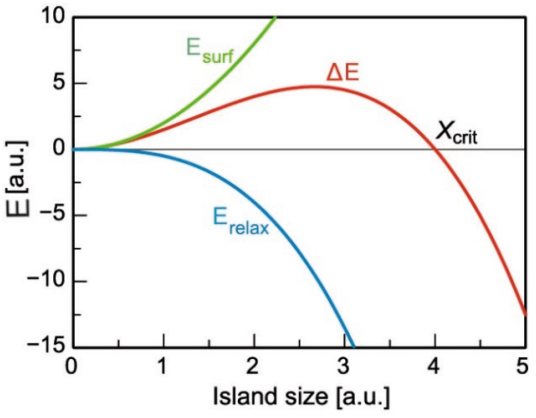


Fig. 12: *Energy difference between a film of flat two-dimensional morphology and a film morphology consisting of three-dimensional islands. The total energy difference and the contributions surface energy difference and relaxation energy are plotted.*

and growth rate the self-organized growth can be close to equilibrium or in the kinetically limited regime. At close to equilibrium conditions (i.e. at high growth temperatures or low deposition flux), the occurring morphology (strained layer or 3D islands or a film with dislocations) is only determined by the energies of the particular configurations. The morphology with the lowest energy will be formed. If the growth is kinetically limited, the activation barriers are important. For instance an initially flat strained layer can transform under kinetically limited conditions to a morphology with 3D islands or to a film with dislocations. What actually happens depends on the kinetics of the growth process, i.e. on the activation energy for formation of 3D islands compared to the activation energy for the introduction of misfit dislocations.

3.3 Semiconductor Nanoislands and Nanowires

Stranski–Krastanov Growth of Nanoislands

Stranski–Krastanov growth occurs for instance in InAs/GaAs growth [6]. An example of InAs nanoislands grown on a GaAs substrate is shown in the TEM image in Fig. 13. The InAs islands were grown by MBE at a growth temperature of 775 K. The density of the islands is $4.5 \times 10^{10} cm^{-2}$ and the lateral island size is 17.5 ± 0.5 nm. The challenges in the growth of these semiconductor islands are to grow islands of desired size and density and with a high size uniformity. As in the case of the 2D islands a higher growth temperature generally leads to the formation of larger islands, a higher growth rate leads to the formation of smaller islands. The size of the islands increases with coverage. Often the density of the islands saturates in an early stage of the growth. These are general trends; details depend on the material system and the particular deposition technique. In some cases (self-limiting growth) the size of the islands saturates and the density increases with coverage. This kind of growth mode leads to a high size uniformity of the islands. The size uniformity achieved in self-organized growth of semiconductor islands can be as small as a few percent. The confinement of charge carriers in nanoscale islands in all three directions gives rise to atomic like energy levels. Quantum dot lasers operating at room temperature have now been realized [11]. The islands grown on a flat substrate are usually not ordered laterally due to the random nature of the nucleation process. In the following it will be shown how nucleation at specific sites can be achieved.

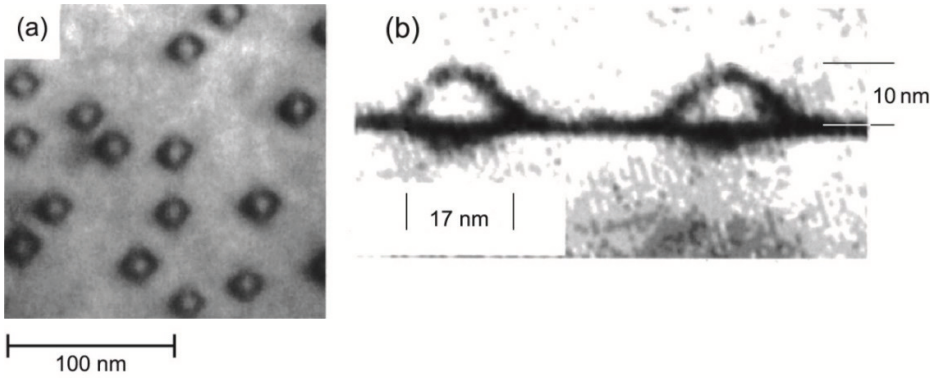


Fig. 13: *InAs nanoislands grown on a GaAs surface. (a) Imaged by plan-view TEM and (b) by cross sectional view TEM [6].*

Lateral Positioning of Nanoislands by Growth on Templates

An example of ordered nucleation at a pre-structured substrate is shown in Fig. 14 [12]. Here Ge islands nucleate above dislocation lines. When a SiGe film is grown on a Si(001) substrate, dislocations form at the interface between the SiGe film and the substrate. The driving force for the formation of the dislocations is the relief of elastic strain which arises due to the different lattice constants between the Si substrate and a Ge/Si film on this substrate. During annealing the dislocations form a relatively regular network, due to a repulsive elastic interaction between the dislocations. The preferred nucleation of Ge islands above the dislocation lines (Fig. 14a) can be explained by local stress relaxation above the dislocation lines providing a lattice constant closer to the Ge one. The nucleation does not occur randomly at the surface, but nucleation occurs simultaneously at sites which have the same structure. This leads to a more narrow size distribution than that for the growth on unstructured Si(001) substrates (Fig. 14b).

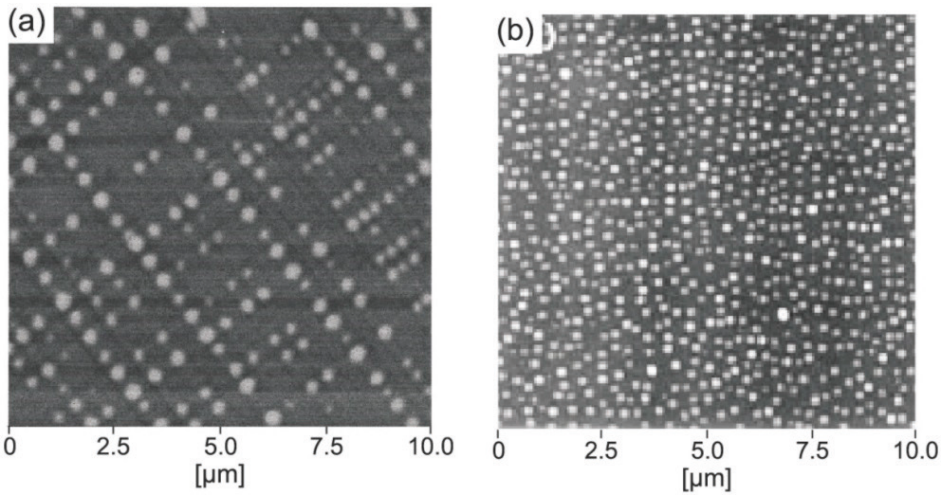


Fig. 14: *(a) Ordered nucleation of Ge islands on a template which is pre-structured by an underlying network of dislocations. (b) Germanium islands grown on a substrate without dislocations [12]. Image sizes 10 µm.*

Monolayer Thick Wires at Step Edges

Monolayer high steps of the substrate surface can be used to fabricate Ge nanowires using step flow growth. Pre-existing step edges on the Si(111) surface are used as templates for the growth of two-dimensional Ge wires at the step edges. When the diffusion of the deposited atoms is sufficient to reach the step edges, these deposited atoms are incorporated exclusively at the step edges and the growth proceeds by a homogenous advancement of the steps (step flow growth mode [5]). If small amounts of Ge are deposited, the steps advance only some nanometres and narrow Ge wires can be grown.

A key issue for the controlled fabrication of nanostructures consisting of different materials is a method of characterization which can distinguish between the different materials on the nanoscale. If the Si(111) surface is terminated with a monolayer of Bi it is possible to distinguish between Si and Ge areas by their apparent height in STM images [13]. Bi, which was deposited initially, always floats on top of the growing layer. Fig. 15a shows an STM image after repeated alternating deposition of 0.15 atomic layers of Ge and Si, respectively. Due to the step flow growth Ge and Si wires are formed at

the advancing step edge. Both elements can be easily distinguished by the apparent heights in the STM images: the height measured by the STM is higher on areas consisting of Ge (red stripes) than on areas consisting of Si (yellow stripes). The apparent height of Ge areas is ~ 0.1 nm higher than the apparent height of Si wires (Fig. 15b). The cross section of a 3.3 nm wide Ge nanowire contains only ~ 20 atoms (Fig. 15c). The width of the wires can be easily tuned by different amounts of Ge and Si being deposited. In this way single-digit nanometre wide nanowire arrays can be fabricated [13].

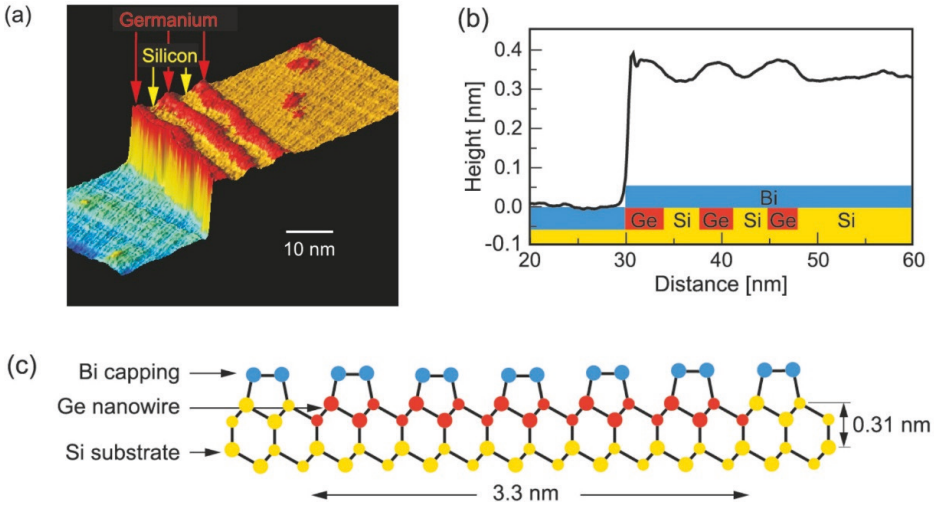


Fig. 15: *(a) STM image of two-dimensional Ge/Si nanowires grown by step-flow at a pre-existing step edge on a Si(111) substrate. Si wires (yellow) and Ge wires (red) can be distinguished by different apparent heights. (b) The cross section across the nanowires. (c) Atomic structure of a Ge wire on the Si substrate capped by Bi. The cross section of the Ge wire contains only ~ 20 Ge atoms [13].*

4 Vapour Liquid-Solid Growth (VLS)

In the VLS growth method [14] small nanoparticles (often gold) induce the growth of nanowires which grow with a diameter determined by the formed by heating a thin deposited gold film which breaks up and reshapes into nanoscale droplets, as shown in Fig. 16a-b. Alternatively gold nanoparticles prepared by wet chemical techniques with sizes of a few tens of nanometers, dispersed onto the surface are used. The melting point of the gold particle is lowered by forming a eutectic mixture with the substrate or the growing species. The actual growth of the nanowires can be performed using different kinds of deposition techniques like molecular beam epitaxy (MBE), chemical vapor deposition (CVD), or metalorganic vapor phase epitaxy (MOVPE). In VLS growth, the growth at the substrate surface is strongly suppressed compared to the growth below the gold particle, which leads to the nanowire growth. The growing species incorporates directly into the gold droplet, or diffuses to the gold droplet and dissolves to the liquid gold. When the concentration of the growing species in the gold droplet exceeds the solubility limit, growth at the solid-liquid interface proceeds, and ideally single crystal nanowires form. The nanowires grow free standing on the substrate and grow (mostly) vertical, as shown schematically in Fig. 16c.

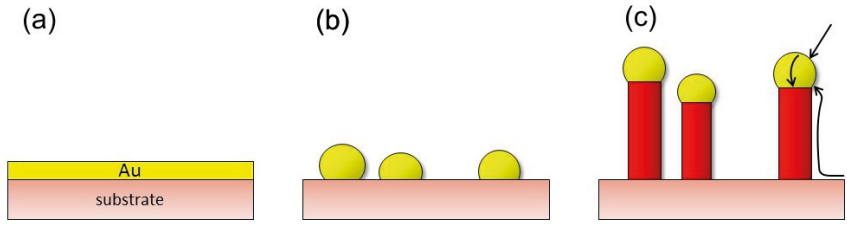


Fig. 16: *Vapor liquid solid growth process. (a) A thin gold film is deposited on the substrate. (b) Subsequent heating of the film leads to decomposition and formation of gold droplets. (c) Incorporation of the growing species into the gold droplet and growth at the liquid-solid interface leads to the formation of nanowires.*

The VLS growth method is used for many different material systems. If the growth at the surface shell of the nanowire is largely suppressed over the growth at the liquid solid interface, the diameter of the nanowires corresponds to the diameter of the gold particle and no tapering occurs. Fig. 17 shows GaAs nanowires of constant diameter which were grown using the VLS method.

The VLS growth method can also be used to grow semiconductor heterostructures within the nanowires with atomically sharp interfaces. Fig. 18 shows an InAs nanowire (green) with several InP barriers (red). The rapid alternation of the composition is controlled by the supply of precursor species from molecular beams to the eutectic melt. Of particular interest is the fact that the very small diameter of the nanowires allows efficient lateral relaxation of lattice constant of the material in the nanowire, thereby providing freedom to combine materials with very different lattice constant to create heterostructures along the nanowire. The problem of incorporation of misfit dislocations when a critical thickness is exceeded does not occur due to the small lateral size of the nanowires.

Fig. 17: *GaAs nanowires grown using the VLS method having a diameter of about 100 nm and a length of 10µm on a GaAs substrate. The three tips of a multi tip STM seen in the image are used to perform four point measurements of the electrical conductivity of freestanding nanowires [15].*

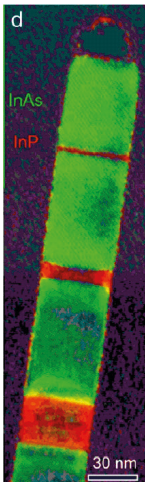


Fig. 18: *(a) Composition profile of an InAs nanowire, containing several InP heterostructures, obtained using reciprocal space analysis of lattice spacings with a TEM. InAs lattice spacings have been color-coded with green and InP spacings with red [16].*

5 Bottom-Up Approaches for Resistive Switching Memories

It has been shown that the VLS growth method can be applied to material systems relevant for resistive switching memories and phase change memories. Figure 19 shows an SEM image of ZnO nanowires, grown by the VLS method. The nanowires are still attached to the substrate. TEM images confirm that these nanowires are single crystalline. Subsequently, the nanowires are removed from the growth substrate and dispersed on a SiO_2 surface for lithographic contacting. Figure 19b shows a sketch of a single contacted nanowire, while in Fig. 19d an SEM image of a contacted nanowire is shown. After an initial "forming" process the two terminal I-V curve shown in Fig. 19c is observed. Initially the device is in the "off" state (red curve), i.e. the resistance is about 10^9 Ω. If the applied voltage approaches 3 V the "Set" switching from the low conductivity "off" state to the high conductivity "on" state with a resistance of about 10^2 – 10^3 Ω occurs. The "Reset" of the device occurs at a reverse voltage of about -1.6 V. Thus a resistive switching memory device with a resistance ratio between on and off states of 10^6 was fabricated. The bipolar switching characteristics indicates that the mechanism of the resistive switching might be the formation/annihilation of metallic Cu filaments originating from the Cu electrode. This assumption is confirmed by control experiments with two Au electrodes, which did not sow resistive switching. Since the nanowires are single crystalline, the widely known formation of conductive filaments at grain boundaries in thin films does not apply for the nanowires. It is assumed that Cu ions may drift along the nanowire surface and form a metal filament outside.

An advantage of nanowire based devices compared to the conventional thin film capacitor-type resistive switching devices is that much more detailed characterization methods can be applied in order to study the switching properties. In the capacitor-type devices only two point conductivity measurements are possible using relatively large contact pads (~ 30 – 100 µm),

as shown in Fig. 20. Devices fabricated from horizontally arranged nanowires allow for (a) multi-probe measurements, (b) measurements under the influence of different atmospheres, and (c) field effect transistor-type measurements, which allow to study the underlying switching mechanisms in much more detail.

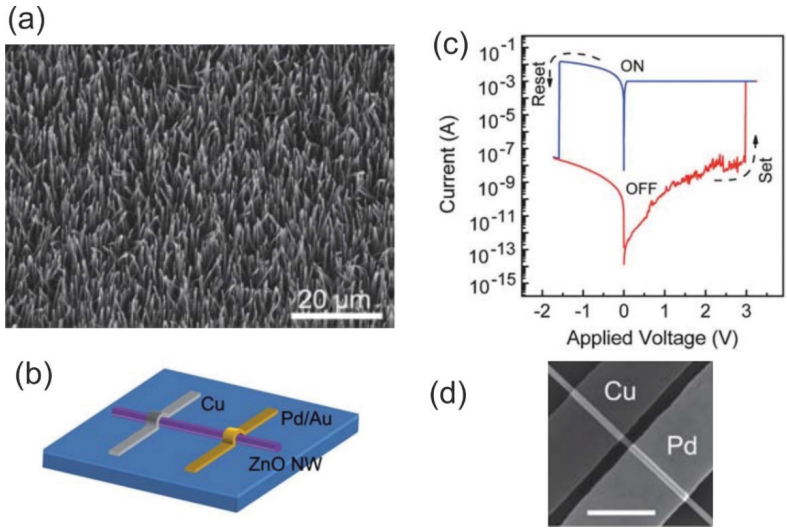


Fig. 19: *ZnO nanowires grown using the VLS method. (a) Dense array of nanowires on the substrate. (b) Sketch of a single nanowire dispersed on a SiO_2 substrate and contacted lithographically. (c) I-V characteristics of the nanowire shown in (d), showing resistive switching characteristics. (d) SEM image of a ZnO nanowire contacted with contact pads. The scale bar is 1µm [17].*

In the nanowire device shown in Fig. 19d the one dimensional character of the nanowire was more a disadvantage than an advantage. It was found that for longer nanowire segments the switching behavior becomes worse. Due to this only a short segment of the nanowire was used as active device component as shown in Fig. 19d. In nanowires known as core/shell nanowires the switching occurs in the radial, not in the axial direction.

The formation of a core/shell nanowire is a two-step process. In a first step a nanowire is grown under the normal VLS growth conditions. Subsequently, in a second growth step the growth conditions are modified (e.g. by changing the growth temperature) such that the growth does not continue in the VLS mode, but that the growth occurs directly at the nanowire shell. Usually the nanowire shell is grown using a different material, leading to a heterojunction between the nanowire core and the shell.

Figure 21a shows a schematic of a Si/amorphous-Si core shell nanowire. In this case core and shell consist of the same material, however, in different phases, i.e. crystalline and amorphous, respectively. The TEM image in Fig. 21b shows the crystalline Si core and the ~5 nm thick amorphous Si shell, which was grown at a lower temperature. The nanowire was lithographically contacted with a Ag electrode shown in Fig. 21a and a second metal (Ni) contact (not shown). The I-V curve shown in Fig. 21c shows a bipolar resistive switching behavior with an onset voltage to the high conductivity state of ~3 V (Ag electrode positively biased).

The switching mechanism is attributed to silver islands that form a conducting filament in the "ON" state.

One can imagine that this kind of resistive switching in a core shell nanowire could be extended towards an implementation into a cross bar array of resistive switching devices.

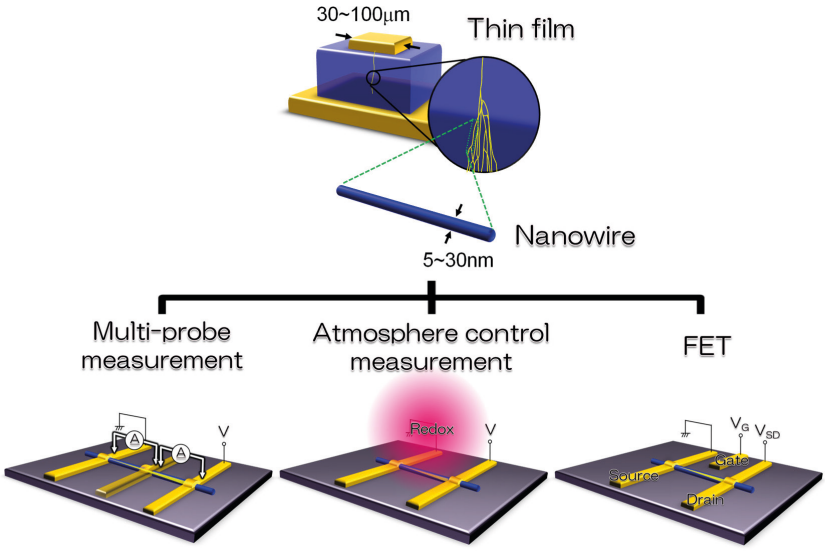


Fig. 20: *Unique features of nanowire resistive switching devices when compared to thin film capacitor-type restive switching devices [18].*

Using the VLS method, nanowires can also be grown from phase change materials leading to the fabrication of a nanowire memory device switching between states attributed to amorphous and crystalline states. Due to the much smaller volume of the material in a nanowire phase change memory, compared to a thin film cell, a reduction of the programming current is expected. Figure 22a shows a single crystalline In_2Se_3 nanowire grown by the VLS method and transferred to a SiO_2 surface. The nanowire was connected by 150 nm wide Pt interconnect lines to outside electrodes.

The device was switched between the low resistance (LRS) crystalline and the high resistance (HRS) amorphous state by voltage pulses of constant width and varying voltage amplitude. After performing the switching, the device resistance was measured at a voltage of 0.2 V. The device is switched from the initial LRS state to the HRS state by pulses with 20 ns width and a very sharp fall down edge (3 ns). The device switches to the HRS state for voltages larger than 7 V, as shown in Fig. 22b. For the opposite switching from the HRS state to the LRS state pulses with a width of 100 μs were applied which lead to a switching event for voltages larger than 5 V.

Once the nanowire is in a specific resistive state, its resistance is stable and the storage of the data, represented by the resistance value, is nonvolatile. The resistance ratio between the two resistive states is 10^5, which is sufficient for nonvolatile memory applications.

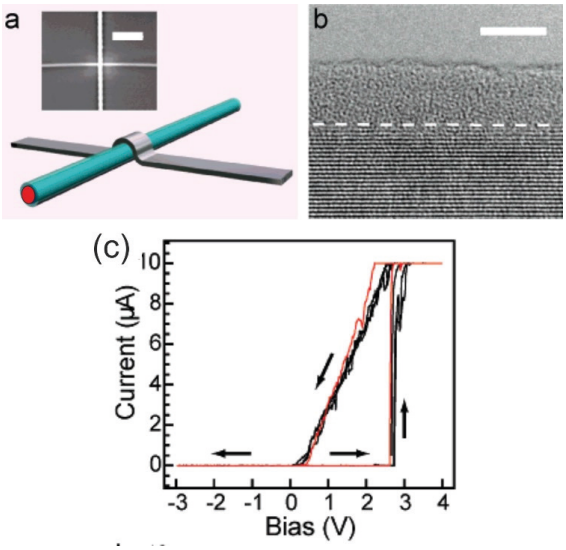


Fig. 21: *(a) Schematic of a core/shell nanowire consisting of a crystalline Si core (red) and an amorphous Si shell (cyan). The nanowire is contacted lithographically by a Ag electrode (gray). The inset shows an SEM image of the actual device. (b) TEM image of the nanowire showing the crystalline core (below the dashed line) and the amorphous shell. (c) I-V curve of the switching characteristic of the device [19].*

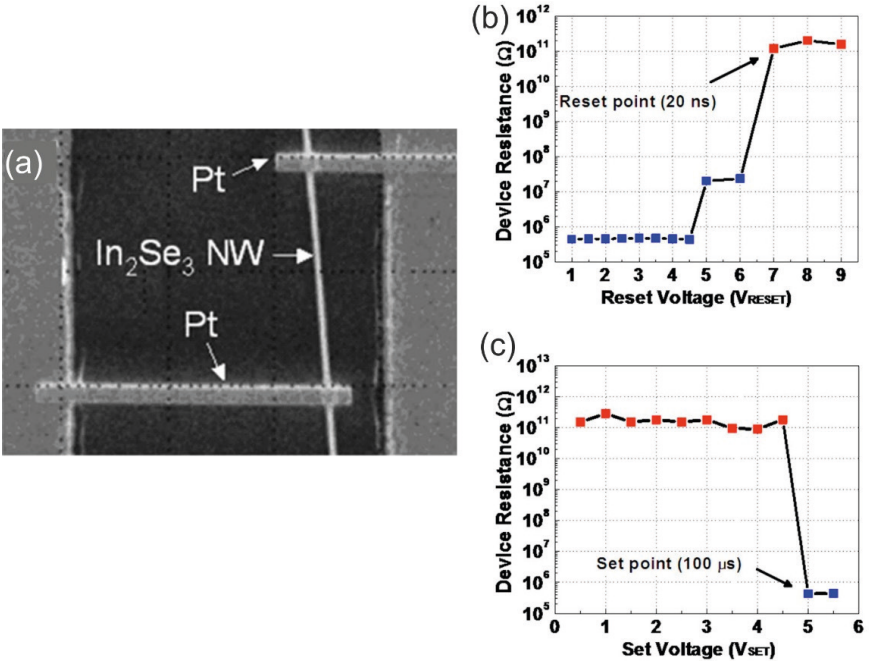


Fig. 22: *(a) SEM image of a In_2Se_3 nanowire grown by the VLS technique, and contacted by metal electrodes. (b)-(c) Device resistance after the application of voltage pulses which were applied in order to switch between the low resistance crystalline and the high resistance amorphous states [20].*

The mechanism of the switching between the two states is due to a phase change between crystalline and amorphous phases through a current induced joule heating process, resulting in the change of electrical resistance. By applying a high and short pulse, the nanowire can be rapidly heated up to its melting point and then quickly quenched. Thus, the nanowire switches to an amorphous phase (high resistance). By applying a low and relatively long pulse, the nanowire is heated up to below its melting point and recrystallizes spontaneously back to a crystalline phase through annealing (low resistance). The dynamic resistive switching ratio for In_2Se_3 nanowires is much larger than that reported for thin film In_2Se_3 devices. Due to its small volume the nanowire memory uses orders of magnitude lower input power and energy to switch between the two material phases. The 20 ns pulse delivers an energy of 1.6 pJ to the nanowire, while the 100 μs pulse delivers only 25 fJ, due to the much higher initial resistance.

References

[1] Nanoelectronics and Information Technology, 3rd Edition, R. Waser (Ed.), Wiley 2012.

[2] L. Vescan, in *Handbook of Thin Film Process Technology*, D. A. Glocker and S. I. Shah, eds., IOP, Bristol, 1995.

[3] E. Kasper, *Silicon Molecular Beam Epitaxy*, Vol. 1-2, CRS Press, 1988.

[4] M. A. Herman, W. Richter, and H. Sitter, *Epitaxy- Physical Principles and Technical Implementation*, Springer, 2004.

[5] B. Voigtländer, Surf. Sci. Rep. **43**, 127 (2001).

[6] V. A. Shchukin, A. A. Lendentsov, and D. Bimberg, *Epitaxy of Nanostructures*, Springer, Heidelberg, 2003.

[7] D. Kandel, Phys. Rev. Lett. **4**, 499 (1997).

[8] S. Filimonov, V. Cherepanov, Yu. Hervieu, and B. Voigtländer, Phys. Rev. B **76**, 035428 (2007).

[9] S. N. Filimonov, Yu. Hervieu, Phys. Rev. E **80**, 051603 (2009).

[10] V. Cherepanov, S. N. Filimonov, J. Mysliveček, B. Voigtländer, Phys. Rev. B **70**, 085401 (2004).

[11] F. Heinrichsdorff, Ch. Ribbat, M. Grundmann, and D. Bimberg, Appl. Phys. Lett. **76**, 556 (2000).

[12] S. Yu. Shiryaev, F. Jensen, J. Lundsgaard Hansen, J. Wulff Petersen, and A. Nylandsted Larsen, Phys. Rev. Lett. **78**, 503 (1997).

[13] M. Kawamura, N. Paul, V. Cherepanov, B. Voigtländer, Phys. Rev. Lett. **91**, 096102 (2003).

[14] S. Barth, F. Hernandez-Ramirez, J. D. Holmes, A. Romano-Rodriguez, Progress in Materials Science **55**, 563 (2010).

[15] S. Korte, M. Steidl, W. Prost, V. Cherepanov, B. Voigtländer, W. Zhao, P. Kleinschmidt, and Th. Hannappel, Appl. Phys. Lett. **103**, 143104 (2013).

[16] M. T. Björk, B. J. Ohlsson, T. Sass, A. I. Persson, C. Thelander, M. H. Magnusson, K. Deppert, L. R. Wallenberg, and L. Samuelson, Nano Lett. **2**, 87 (2002).

[17] Y. Yang, X. Zhang, M. Gao, F. Zeng, W. Zhou, S. Xie and F. Pan, Nanoscale **3**, 1917 (2011).

[18] Resistive Switching, I. Daniele and R. Waser (Ed.), Wiley 2016.

[19] Y. Dong, G. Yu, M. C. McAlpine, W. Lu, and C. M. Lieber, Nano Lett. **8**, 386 (2008).

[20] B. Yu, S. Ju, X. Sun, G. Ng, T. D. Nguyen, M. Meyyappan, and D. B. Janes, Appl. Phys. Lett. **91**, 133119 (2007).

C 1 Electrical Characterization of Memristive Cells

U. Böttger and V. Havel
Institut für Werkstoffe der Elektrotechnik 2
RWTH Aachen, Sommerfeldstr. 24, 52074 Aachen

Contents

1 Introduction **2**

2 Quasi-static current-voltage measurements **3**
2.1 Measuring principle 3
2.2 Electroforming 4
2.3 Dependence on sweep rate 5

3 Current compliance and overshoot effects **6**

4 Pulse measurements for study of switching dynamics **9**
4.1 Basic experimental setup 9
4.2 Comparison of quasi-static and pulse measurement 11
4.3 Experimental setup and results for nanosecond switching 12

5 Pulse measurements for sub-nanosecond switching **13**
5.1 High-frequency basics 13
5.2 HF measuring setup 14
5.3 Design of integrated resistive switching cell 16
5.4 Ultra-fast resistive switching of tantalum oxide cell 17

6 Conclusion **19**

Lecture Notes of the 47^{th} IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

Resistive switching describes a variety of phenomena where the resistance of a two-terminal element changes under an external electrical field between two stable states, a high-resistive state (HRS) and a low-resistive state (LRS), e. g. [1]. Using a suitable writing procedure, multiple resistance states are possible which may enhance the number of potential applications. Due to the hysteretic behavior of the resistance changes, this property has been called memristive firstly introduced by Leon Chua [2]. Hereon based memories are referred to as RRAM (resistive random access memory) - also ReRAM, OXRAM, CBRAM, etc. when a specific functionality or feature should be emphasized. From the device point of view one of the most interesting aspects of RRAM is the dynamic characteristic. Unlike transistors, diodes, resistors and the majority of electronic devices, resistive switching devices change their intrinsic properties by external stimuli depending on their strength and duration (non-linearity and time-variance).

The exact physical mechanisms determining the switching process are still under discussion: trap controlled space charges [3], Schottky barrier height changes [4], diffusion of oxygen vacancies and associated bulk and interface contributions [5], formation and dissolution of metallic filaments [6] are effects among others which might play a crucial role.

Standard characterization of RRAM is mostly undertaken by quasi-static techniques available by commercial, conventional semiconductor parameter analyzers. Quasi-static investigations are sufficient for a first assessment of memory cells in the framework of material and process optimization, however, possibly critical time-dependent effects may be overlooked. A crucial property for resistive switching is the speed of the transition between the states involved. To understand this phenomenon, it is necessary to determine the mechanisms related to switching kinetics by measuring switching processes on a broad timescale between sub-nanoseconds and days. Most reports published so far about the microscopic mechanisms of resistive switching effects for non-volatile memory applications do not take into account, that a significant voltage-time dilemma needs to be overcome for the use in memory cells [7]. One needs a mechanism which changes the state of the memory by a write voltage pulse of several ns, while it must withstand read voltages for up to 10 years, i. e. 3×10^8 s. Nevertheless, the read voltage may not be less than approximately a tenth of the write voltage. Obviously, there must be switching kinetics involved with an extremely high degree of non-linearity. Within one order of magnitude on the voltage scale, the switching time needs to change by not less than 16 orders of magnitude.

Impedance spectroscopy is a suitable tool to study defects and trapping phenomena of deep level states in semiconductors [8]. The measurements deliver the real and imaginary parts of the dielectric constant. From those characteristic times are derived that are linked to specific polarization processes as electronic, ionic, dielectric or others. Based on such investigations Menke et al. [9] developed a microscopic model for different resistance states and the nature of the conducting channels for resistive switching.

In the framework of this contribution, two approaches of investigating the device dynamics are presented: (i) a current compliance (CC) during of quasi-static I-V measurements controls the runaway current in the event of switching at which CC is realized by internal features of the measurement equipment or by external devices (e. g. field-effect transistors), (ii) an ultra-fast stimulus and detection is applied in order to "resolve" the process of resistance change in the best possible way.

The paper is based loosely on Chapter 12 "Quasi-static and Pulse Measuring Techniques" of Torrezan, Medeiros-Ribeiro, Tiedke which is part of book "Resistive Switching" [10].

2 Quasi-static current-voltage measurements

2.1 Measuring principle

The current-voltage characteristic of a resistive-switching random access memory cell is the simplest way to evaluate its functionality. The measurement setup consists of (i) a controllable voltage source and an amperemeter or (ii) a current generator and a voltmeter. In case (i) the voltage across the cell is swept while monitoring the current through the device, in case (ii) the voltage is measured as a function of the current variation. For sufficient slow sweep of the stimulus the method is referred as quasi-static or quasi-DC. The topology for this technique is found commercially in the form of SMUs (Source Measurement Unit) with a range of current, voltage, and power capabilities available [11]. SMUs can also be utilized as stand-alone current and voltage sources as well as a stand-alone ammeter, voltmeter, ohmmeter, and electronic load.

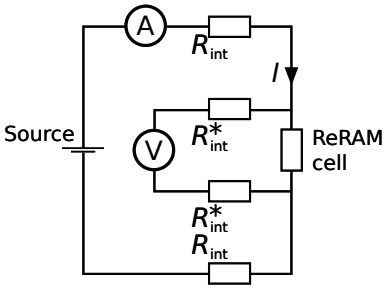


Fig. 1: *Four-terminal sensing driven by a voltage source. R_{int} and R^*_{int} denote the interconnect resistance.*

Four-terminal sensing, also known as four-wire sensing or Kelvin sensing (Fig. 1), comes into play when more precise measurements than the conventional two-terminal sensing are required. Due to the interconnect resistances R_{int} the applied voltage of the source may significantly differ from the voltage drop across the device under test (DUT). If the input impedance of the measuring unit is sufficient high, the voltage drop V is determined currentless and the cell resistance is given by V/I. Since almost no current flows in the voltage sensing pair, interconnect resistances R^*_{int} to the cell bottom electrode (BE) as well as to the top electrode (TE) will not distort the sensing.

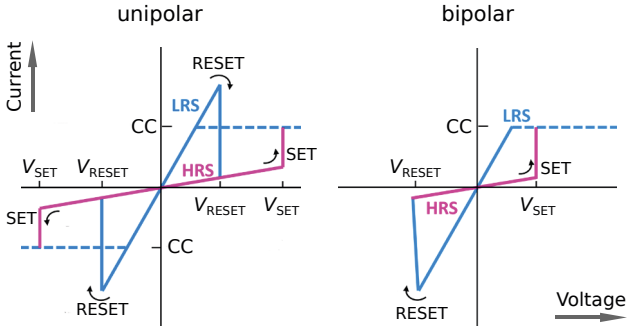


Fig. 2: *Unipolar and bipolar I-V characteristics of resistive switching cells. SET (or ON switching) at V_{SET} is the process from HRS to LRS, RESET (or OFF switching) at V_{RESET} corresponds to the reverse switching from LRS to HRS. CC guarantees irreversible cell breakdown during SET.*

As shown in Fig. 2, quasi-static I-V measurements are able to identify characteristic switching voltages V_{SET} and V_{RESET} in which the high resistance state (HRS) or OFF state passes into the low resistance state (LRS) or ON state and vice versa. If the ON and/or OFF transition is gradual (not shown), V_{SET} and/or V_{RESET} may be specified at a certain percentage decrease (increase) in resistance [12].
Fig. 2 also displays the two principle switching modes of resistive memories: (i) the unipolar and (ii) the bipolar mode. The first-mentioned one is independent of the polarity of the switching voltages, V_{SET} and V_{RESET} are observed for positive as well as for negative values. It is characteristic that ON switching needs higher voltages than the RESET process [13]. The bipolar mode needs both the polarities for full operation. A current compliance is applied in all cases where irreversible cell breakdown during SET have to be avoided and possible multi-level cell switching has to be controlled [14].

2.2 Electroforming

For stable operation with repeatable characteristics, an electroforming procedure (short: forming) of pristine cells is usually needed prior to the resistive switching. The forming behavior is illustrated exemplary for two bipolar switching cells, based on strontium titanate (STO) and tantalum oxide (TaO_x), as step "0" in the I-V graphs of Fig. 3 [15]. A common feature is the fact that the forming voltage is higher than V_{SET} for subsequent switching. However, the two material systems differ in voltage polarity. Where STO necessitates positive voltage and electroforms into the HRS, the situation is vice versa for TaO_x. The rather extraordinary forming process for STO is extensively described in [13], whereas the forming in TaO_x is believed to be more an initial SET process, e. g. [16], not significantly differing from subsequent set processes.

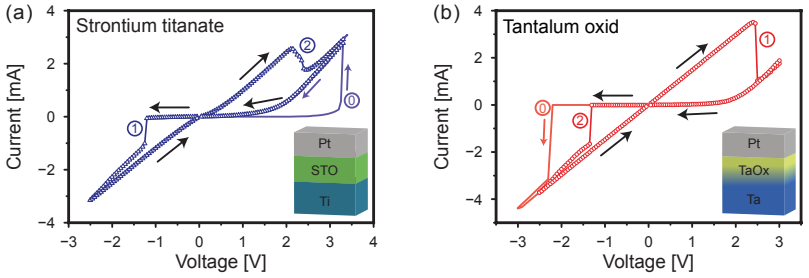


Fig. 3: *Electroforming and switching in material systems (a) Ti/$SrTiO_3$/Pt and (b) Ta/TaO_x/Pt measured by triangular voltage signals. The numbers label the forming process (0), set (1), and reset (2) for STO and TaOx. The voltage is applied to the top electrode with the bottom electrode grounded, the slew rate usually is in the range of volts per second. Adapted from [15].*

I-V measurements will not only deliver switching characteristics but also other information about the conduction behavior in the LRS and HRS. On the one hand this will contribute to identify the physical conduction mechanisms of the resistance state (like Schottky emission, tunneling etc. [17]) and to improve the understanding of the switching process itself; on the other hand the non-linear properties of the resistance states can be analyzed which becomes important to develop specific device strategies to overcome the problem of sneak path leakage current in passive cross-point memory arrays without selectors [18].

2.3 Dependence on sweep rate

The kinetic of the switching process is directly linked to the sweep rate r_V of the triangular voltage signal during the I-V measurements. Fig. 4 a and b show the measured I-V characteristics of ON and OFF switching of HfO_x-based resistive memory devices with V_{SET} and V_{RESET} increasing as the sweep rate is increased [12].

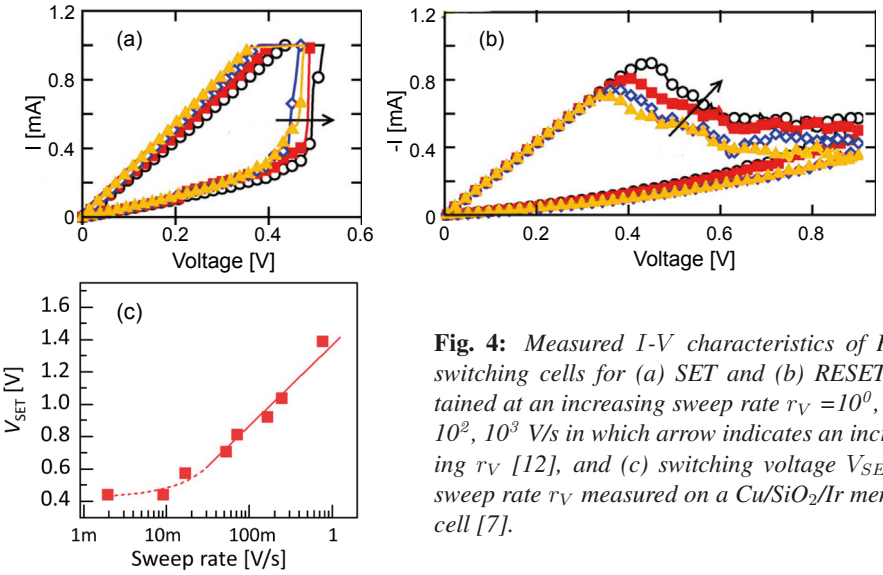


Fig. 4: *Measured I-V characteristics of HfO_x switching cells for (a) SET and (b) RESET obtained at an increasing sweep rate r_V =10^0, 10^1, 10^2, 10^3 V/s in which arrow indicates an increasing r_V [12], and (c) switching voltage V_{SET} vs sweep rate r_V measured on a Cu/SiO_2/Ir memory cell [7].*

Similar trend for V_{SET} are also present in Cu-SiO_2-based resistive cells as illustrated in Fig. 4 c [7]. Since the sweep rate increases, the resistive cell has less time to switch and needs a higher voltage. A clear exponential relationship between the switching voltage and the sweep rate is observed for medium to high sweep rates r_V, while for low r_V values, a critical SET voltage seems to be approached. The pronounced exponential relationship and, in particular, a critical threshold voltage for the SET process explain how the voltage-time dilemma is overcome for the SET process in resistive switching memory cells.

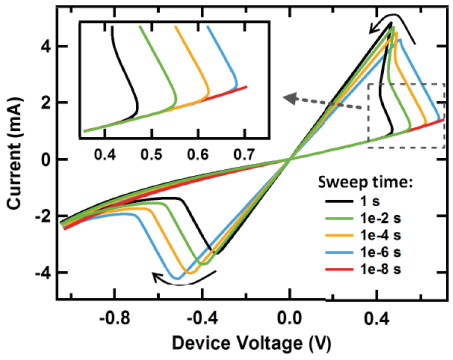


Fig. 5: *Simulated results for current-voltage sweeps with varying voltage ramp rates giving a total sweep time from 10 ns to 1 s. A positive voltage sawtooth to $+0.8$ V is applied followed by a negative sawtooth to -1.2 V. $R_{external} = 70\ \Omega$, and voltage plotted is the internal device voltage.*

The relation between V_{SET} and sweep rate r_V is also important for the device modeling of RRAM cells. In [19] the I-V behavior of TaO_x cells was simulated for different sweep rates. As seen in Fig. 5, V_{SET} and V_{RESET} increase for faster sweep rates. In case of very fast sweep times like 10^{-8} s that is corresponding to the red line in Fig. 5, the applied voltage is too low in order to initiate the ON switching. A collapse of the pinched hysteresis loop to a single trace is present. By the means of quasi-static I-V measurements the dependence of the switching voltage on the sweep rate can be observed, however, the accessible information about the switching dynamics, i. e. the time-dependent evolution of the switching process, is limited, see Sec.“pulse measurements“.

3 Current compliance and overshoot effects

As already mentioned above, the use of current compliance for the voltage source prevents the cell from an irreversible breakdown during electroforming (or more general from an over switching), and controls the maximum current I_{CCmax} for multi-level operation. An example for tuning the device resistance states of a multi-level cell is shown in Fig. 6 [20].

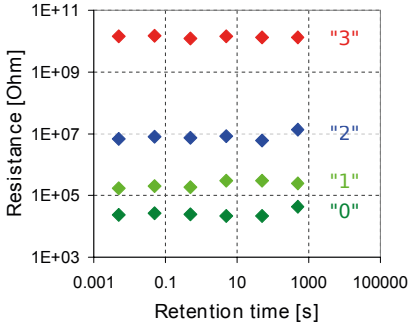


Fig. 6: *Result of a short-term retention test with multi-level states of CBRAM cells. “0“ is representing the OFF state, different ON states were written by different current compliances: $I_{CC} = 20$ nA (“1“), $I_{CC} = 1$ μA (“2“), $I_{CC} = 13$ μA (“3“). The cells are re-written for each measured retention period [20].*

There are two general ways to implement the CC functionality in an electronic circuit: (i) passive and (ii) active current compliance. A passive CC bases on limiting the current with a series resistor (as shown in Fig. 7 a) determining the maximal current for the extreme case of short circuiting the DUT at an applied voltage V_S as $I_{CCmax} = V_S/R_{CC}$.

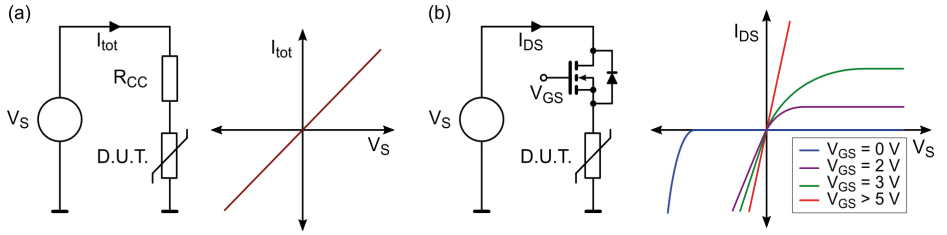


Fig. 7: *Passive current limiting with (a) a series resistor and (b) a series transistor. The third quadrant of the characteristic is influenced by the body diode*

This simple method does not have any major drawbacks and represents the safest current limiting solution. However, it should be guaranteed that the limiting resistor is placed as close as possible to the DUT. In that case, the influence of parasitic capacitances C_{par}, which may originate from scope probes, cables (e. g. $C_{\text{par,BNC}} \approx 1$ pF/cm for a 50 Ω BNC cable), probe needles, or device electrodes, can be neglected. Otherwise, charge and discharge of such a capacitance would bypass the compliance element R_{CC} in the moment of switching [21]. The capacitor would experience a voltage drop at the rapid change of R_{DUT} leading to a signal overshoot at the DUT. In the corresponding equivalent circuit C_{par} is parallel to the source and in series to the DUT, the resistor and the oscilloscope. Fig. 8 of the transient device currents during the forming of NiOx-based cells illustrates the benefit of a favorable arrangement of the setup and the prevention of overshooting.

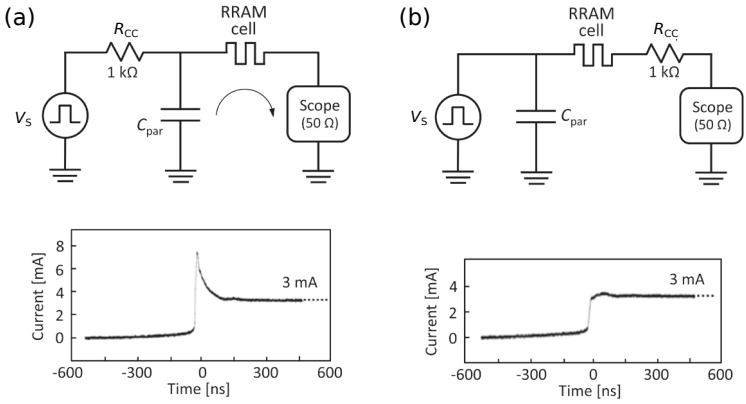


Fig. 8: *TiO_x RRAM cell during electroforming: equivalent circuit with the compliance resistor R_{CC} and parasitic capacitances C_{par} and measured device current for different placings of R_{CC}: (a) far away from the memory cell, and (b) close to the cell . Adapted from [21].*

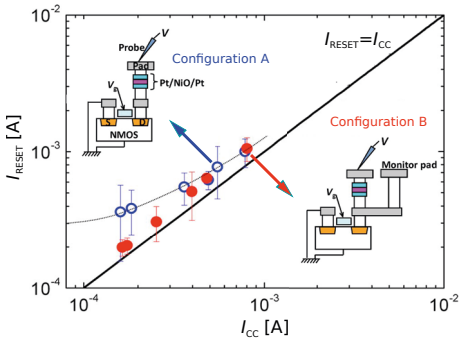


Fig. 9: *Mean measured I_{RESET}-I_{CC} characteristics of successive thirty switching cycles for two different configurations: The circuit composed by connecting (i) the 1T1R cell consisting of the Pt/NiO_x/Pt structure and the cell transistor, Configuration A, and (ii) the 1T1R cell with the monitor pad, Configuration B. Adapted from [22].*

Instead of a resistor, a bipolar or a FET type transistor, connected in series and operating in the saturation (active) regime, can be used as a current restricting element (see Fig. 7 b). For a FET transistor, the current compliance level I_{CC} is adjusted by setting the transistor gate voltage V_{GS} to a corresponding level of saturated drain current (given by the transistor output

characteristics), and not controlled by the actual I_{DUT}, i.e., there is no feedback control. From this point of view the transistor is also a passive CC system. Further, the actual V_{DUT} must be compensated ($V_{\mathrm{DUT}} = V_{\mathrm{S}} - V_{\mathrm{DS}}$). However, V_{DS} is non-linear and, therefore, the compensation is not trivial. Furthermore, a bipolar CC can not be realized by the means of a single transistor (due to the asymmetric output characteristic).

Similar to the resistor, optimal use of the transistor as a compliance element needs minimizing the parasitic capacitance between the RRAM cell and the transistor. During switching V_{DS} increases resulting in a transient current that charges the parasitic capacitance and flows through the DUT without any control by the transistor. For an unipolar switching device the influence of the parasitic capacitance is illustrated by the deviation of two different cell configurations from the expected relation $I_{\mathrm{RESET}} \approx I_{\mathrm{CC}}$ in Fig. 9 due to transient overshoot currents. Configuration A has a higher parasitic capacitance than configuration B resulting in a higher I_{RESET} and in a more pronounced deviation from the expected behavior, especially at low I_{CC} [22].

The active CC bases on the monitoring of I_{DUT} and controlling the SMU sourcing output, which is switched between the voltage source and the current source [23]. The operating principle is shown in the block diagram of Fig. 10. It is obvious that the active CC is a feedback system. The circuit (switch control) cannot react instantaneously since the output current must be detected and compared to the threshold level I_{CC} before. During this response time, which is approximately 60 ns at best, the current can significantly exceed the target compliance value and represents a potential risks for the DUT.

Fig. 11 shows a typical delay in the response of a SMU with integrated current compliance. A significant current overshoot is observed at the beginning of the pulse at 2 s with the current compliance being enforced after a delay of 125 ns.

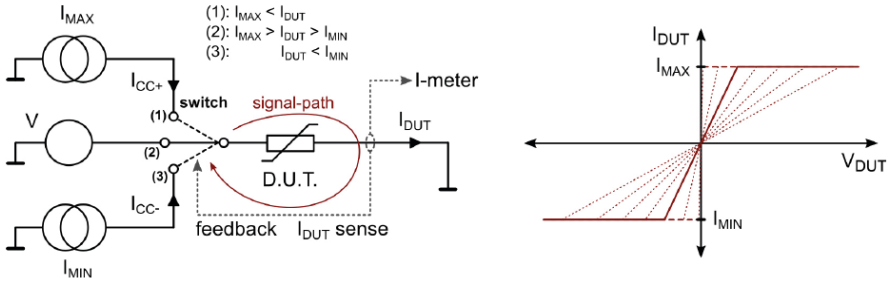


Fig. 10: *Principal schematic and characteristics of SMU active current compliance.*

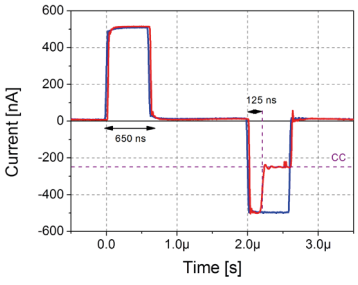


Fig. 11: *Transient device current (red curve) in a NiOx RRAM cell as measured by a SMU with integrated current compliance. Adapted from [10].*

4 Pulse measurements for study of switching dynamics

4.1 Basic experimental setup

The measurement of the time-dependent transient response of voltage and current to a stimulus as a voltage pulse or a voltage step is a suitable tool to investigate switching dynamics of ReRAM cells and to identify memory characteristics such as switching time and switching energy. The minimum observable timescale is determined by the slowest component of the setup: rise and fall times of the voltage driver, the bandwidth of the measuring system to track the switching voltage and current in real time, or intrinsic RC times from parasitic resistance and capacitance associated with interconnects and the device structure.
A typical setup for the pulse measurement is illustrated in Fig. 12. A rectangular voltage pulse from a generator was applied to the cell, and the voltage drop across the device is monitored by an oscilloscope. The corresponding transient current detection by a further channel of the oscilloscope utilizes its 50 Ω input impedance, a combination of an external resistor and an amplifier, or another method fulfilling the target bandwidth. If there is a need to protect the device against breakdown as already discussed in Sec. 3, a current limiting element such as a series resistor or a transistor has to be added.

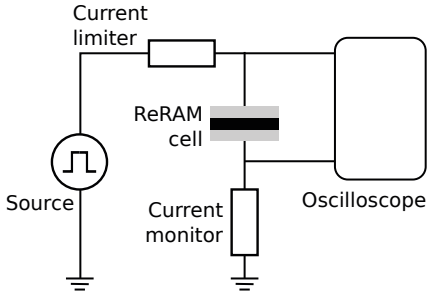


Fig. 12: *Typical setup for the pulse measurement to investigate ReRAM device dynamics. The location of the current limiting element may differ from the one in the illustration.*

As an example, Fig. 13 a shows the results of a Ti/STO:Mn/Pt stack with a sputtered 25 nm STO film on a 50 nm thick Ti bottom electrode. Top electrodes of the MIM cell are realized by 100 nm Pt film patterned into 50 × 50 μm^2 pads. The voltage and current transients were measured using a Keithley 4225 PMU (pulse measurement unit) with two remote amplifiers when a rectangular voltage pulse for SET is applied to the cell [24]. The measurement cycle begins with +4.5 V RESET for a 10 μs, followed by a READ at 0.5 V for 1 ms. The subsequent SET pulse is varied between -1 V to -3 V and 1 μs up to 0.9 s. A second READ verifies the SET and completes the cycle (inset Fig. 13 b). For an applied voltage of -2.5 V a pulse width $t_{\mathrm{SET}} = 50\,\mu$s was found. When the voltage rise a charging current of cell capacitance with $R_{\mathrm{ext}}C = 72$ ns is observed followed by an exponential relaxation. The relaxed current remains on a constant level of 150 μA for 30 μs before it jumps to a higher level of 300 μA. The levels are identified as the HRS and the LRS. The SET time t_{SET} is the time between the beginning of the pulse and the onset of current increase. Fig. 13 b illustrates by plotting "$\log t_{\mathrm{SET}}(V_{\mathrm{SET}})$" that the switching time steeply decreases as a higher voltage is applied. Between -1.35 V$\leq V \leq -2.5$ V the SET time shows a strong non-linearity of approximately seven orders of magnitude (1 s to 200 ns). In further references about studies of switching kinetics in STO-based ReRAM devices even nine orders of magnitude was found for an icrease of the voltage

from 1 V to 5 V [5]. In HfOx/AlOx bilayer cells a decrease of four orders of magnitude in the pulse width is achieved when the pulse height is increased by about 1 V [25].

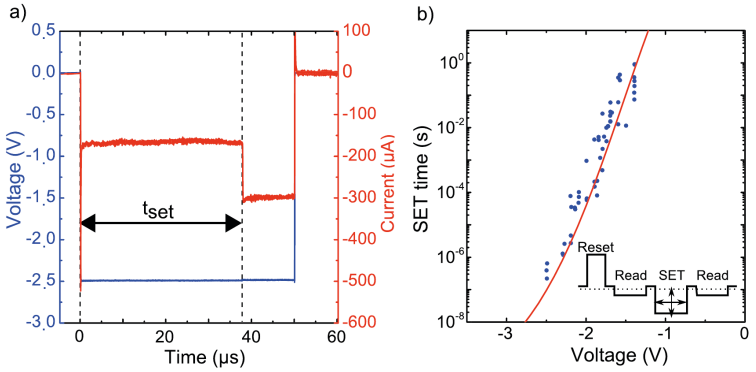


Fig. 13: *(a) One exemplary voltage (blue) and current (red) transients of switching of a STO memory cell after applying a 2.5 V SET pulse, (b) t_{SET} as a function of the applied voltage, experiment (blue) vs. model (red). The inset shows the applied pulse train schematics, see also [24].*

From the measured device voltage and device current it is possible to extract many other dynamical variables for the SET as well as for the RESET process: device conductance G, resistance$R$, energy differential of conductance dG/dE etc, see [26]. An example is shown in Fig. 14: G changes considerably with the initial energy expenditure, but after that there is barely no change with any additional energy injected resulting in a waste of energy and/or a potential cell damage. Due to the fact that a steep decrease in switching energy (analogue to the switching time mentioned above) with increasing the applied voltage reduced write energy can be obtained by using short write pulses with higher voltage amplitude. [25, 27].

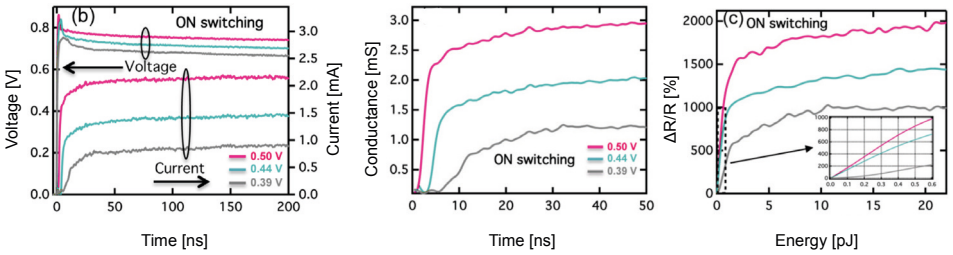


Fig. 14: *(a) ON-Switching of a 10 × 10 µm^2 TaO_x cell by three different positive amplitude pulses applied at time t = 0 for 300 ns, and (b) conductance as function of time, (c) resistance change in per cent $(R_{OFF} - R_{ON})/R_{ON}$ as a function of energy expended during ON switching. [26].*

4.2 Comparison of quasi-static and pulse measurement

As already illustrated in Fig. 13 a and shown again in Fig. 15 a the switching time t_{SET} can directly be determined from pulse measurements. The same information is achieved by the means of I-V sweep measurements with a given sweep rate r_{V}. Using the relation

$$t^*_{\mathrm{SET}} = \frac{V_{\mathrm{SET}}}{r_{\mathrm{V}}} \qquad r_{\mathrm{V}} = \frac{\Delta V}{\Delta t} \tag{1}$$

a switching time t^*_{SET} can be calculated from the sweep measurements (Fig. 15 b). It is obvious that $t^*_{\mathrm{SET}} = t_{\mathrm{SET}}$. A common plot of switching time data in Fig. 15 c confirms the conclusion. The reverse procedure is also possible: an I-V curve can be constructed when the current at a given voltage is taken after a specific time Δt_0 from pulse measurements of Fig. 15 a. After Δt_0 the current may be in the HRS, LRS or in an intermediate state. For $\Delta t_0 = 1$ ms

$$I(V)|_{\Delta t_0 = 1\,\mathrm{ms}} \tag{2}$$

shows a qualitatively identical behavior compared with the direct measured I-V curve, see Fig. 15 d. The different switching voltages results from the uncertain assignment of the voltage ramps.

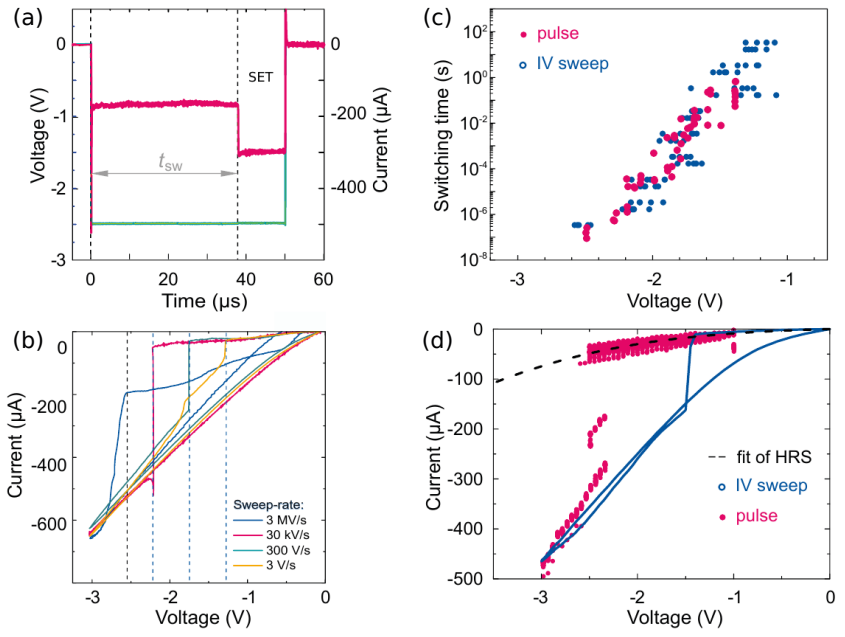


Fig. 15: *(a) Transient voltage and current response , (b) I-V sweeps with different voltage ramps r_{V} in which only the quadrant of the SET process is shown, (c) switching times by different techniques, and (d) directly measured and (from the results of pulse experiments) constructed I-V sweeps.*

4.3 Experimental setup and results for nanosecond switching

For the observation of voltage and current dynamics in RRAM cells in the nanosecond timescale, the technique shown in Fig. 16 has been used [28, 29]. Pulses of 1 ns at full width at half maximum (FWHM) are passing an impedance matching network which produces an attenuated signal in the pick-off port and another with little attenuation towards the electrode of the device. The other device electrode is connected to R_{shunt} monitoring the device current through two stages of operational amplifiers (OPA) with an overall voltage-to-voltage gain of 10x and 200x, respectively. The signals were sampled by a fast oscilloscope with GHz bandwidth. Besides minimizing parasitic capacitance, any remaining capacitor in the sample path should see a low equivalent resistor wherever is possible to reduce RC loading.

The setup includes an AC method to determine the device state before and after a fast pulse. It is comprised of a lock-in amplifier, located at the 400x output of the second stage of the OPAs and a kHz sin wave generator with an amplitude small enough to avoid any disturbing of the memory cell state. Here, device resistance up to 50 MΩ can be measured.

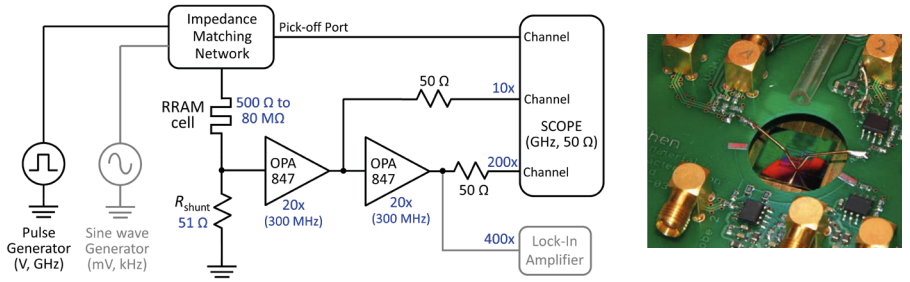


Fig. 16: *Combined setup for the investigation of switching dynamics in RRAM devices in the nanosecond regime: (a) scheme and (b) picture of the PCB with access the cell by probing needles [28].*

As an example nanosecond switching of nanocrossbar TiO_2 devices was demonstrated Fig. 17. The observed high peak current could cause a high cell temperature by Joule heating, which may able to accelerate the switching process, i. e. to reduce the switching time [29]. Switching time of 1 ns has also been observed in GeTe phase change memory cells utilizing the described setup [30].

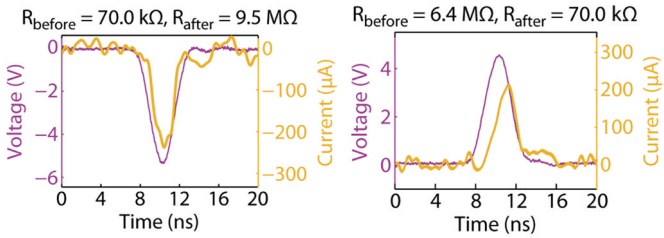


Fig. 17: *(a) RESET and (b) SET for 100 × 100 nm^2 nanocrossbar TiO_2 cell [29].*

5 Pulse measurements for sub-nanosecond switching

5.1 High-frequency basics

Due to the fact that signal processing in the high frequency (HF) range has to be described rather by the point of view of wave propagation than by conventional voltage and current, ultra-fast transient electric measuring techniques require a specific approach. At the upper MHz, GHz and over-GHz frequency range the components and interconnections are considered as microwave (MW) components and transmission lines (TLs), i.e. as elements with distributed parameters instead of lumped elements. Voltage and current vary in magnitude and phase over the length of the MW component or the TL. Critical values for the adoption of the TL concept can be given in the time regime [31] as well as in the frequency/wavelength regime [32]:

$$\Delta t \leq \frac{1}{2}\, t_{\text{rise}} \qquad\qquad \ell \leq \frac{\lambda}{8} \tag{3}$$

The first expression benchmarks the one-way propagation delay Δt due to the lengths of cables, wires or PCB (printed circuit board) tracks with respect to the applied signal rise time t_{rise}, the second one compares the physical length of the signal line ℓ with one eighth of the wavelength of the signal λ. Assuming a signal rise time $t_{\text{rise}} = 500$ ps and a relative dielectric constant of $\varepsilon_{\text{r}} = 4.0$, a critical TL length $\ell > c_0 t_{\text{rise}}/2\sqrt{\varepsilon_{\text{r}}} \approx 7.5$ cm is calculated from the first part of Eq. 3.

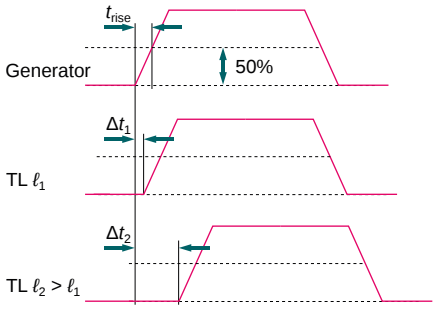


Fig. 18: *Propagation delay Δt due to the path lengths with respect to the applied signal rise time t_{rise}.*

In addition, the rise and fall times t_{rise} and t_{fall} determine the minimal bandwidth of a system required for a correct signal processing. An approximate rule is expressed by

$$f_{\text{max}} = \frac{0.35}{t_{\text{rise}}, t_{\text{fall}}}. \tag{4}$$

In order to guarantee an optimal power transfer from the source to the load or any interconnection in the HF system, impedance matching of load and source resistance is a crucial requirement (”maximum power transfer theorem” [33]). A TL is modeled as an infinite series of elementary components specified per unit length as a series resistor R', series inductor L', shunt capacitor C', and a shunt conductance G' resulting in a TL impedance:

$$Z_0 = \sqrt{\frac{R' + j\omega L'}{G' + j\omega C'}} \tag{5}$$

The majority of HF components features a characteristic impedance $Z_0 = 50\ \Omega$. In case of impedances mismatch of a device under test (DUT) or another component or junction with

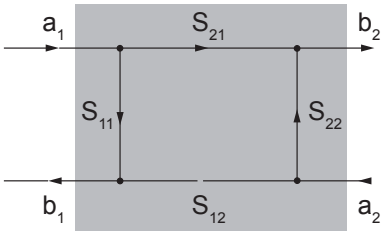


Fig. 19: *S-parameter of a 2-port network with incident power wave* $a_i = \left(V_i/\sqrt{Z_0} + Ii\sqrt{Z_0}\right)/2$ *and transmitted or reflected power wave* $b_i = \left(V_i/\sqrt{Z_0} - Ii\sqrt{Z_0}\right)/2$ *which are defined as linear combinmations of voltages and currents.*

$Z_1 \neq Z_0$ in the line, a part of the power will be reflected back to the source. The rate of reflection is described by the reflection coefficient:

$$\Gamma = \frac{Z_1 - Z_0}{Z1 + Z_0}. \tag{6}$$

Each individual component whose impedance matching is not ideal can be regarded as a two port network element. Typically the scattering (S-)parameters are used in the HF regime to characterize such a two port network, e.g. [34]. The S-parameters relate to relatively simple measurements of power waves such as gain, loss, and reflection coefficient and avoid connections of undesirable loads to the DUT [35]. Fig. 19 illustrates the conventions of the incident and reflected/transmitted power wave direction and the parameter flow on a generic two-port network using the two-dimensional relations between the scattering matrix and the power waves. a_i represents an incident wave, b_i a reflected or transmitted one:

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \qquad S_{ij} = \left.\frac{b_i}{a_j}\right|_{a_k=0}, \; k \neq j. \tag{7}$$

S_{11} and S_{22} are the forward and reverse reflection coefficients with the opposite port terminated by $Z_0 = 50\ \Omega$, and S_{12} and S_{21} are the forward and reverse gains assuming a termination at both the source and load site with $Z_0 = 50\ \Omega$ in order to fulfill the measurement instruction $a_k = 0$ as defined in (7). In the framework of characterization of resistive switching devices S-parameter are commonly used for frequency-dependent network analysis and optimization of circuit design with respect to impedance matching. Besides the optimization of lines, cables, interconnects, the most critical component of the HF test setup is the resistive switching cell itself. Its variable resistance - which is just the desired functionality - turns it into an imperfectly matched component.

5.2 HF measuring setup

A scheme of a suitable testing setup for the observation of ultra-fast switching with real time monitoring is shown in Fig. 20 [37]. A fast source generates a single-shot, short pulse which is split into a low-loss TL comprising the DUT, i.e. the ReRAM cell, and into a reference TL. Both the lines are attenuated and routed into the ports of a high-bandwidth, multi-channel oscilloscope terminated by an impedance $Z_0 = 50\ \Omega$. All the interconnections in the system are generally considered as ideally 50 Ω impedance matched with exception of the DUT. Therefore, the applied signal is partially reflected at the mismatched interconnection. As already mentioned above, the DUT even changes its resistance during the pulse. Fig. 20 visualizes the influence of an ideal series resistor R_{DUT} to the transient signal behavior. It is obvious that the scope

monitors not only the transmitted voltage pulse V_{trans} through the ReRAM (green path), but also the incident voltage pulse V_{inc} from the pulse generator (red path) and the reflected voltage pulse V_{ref} from the DUT (blue path).

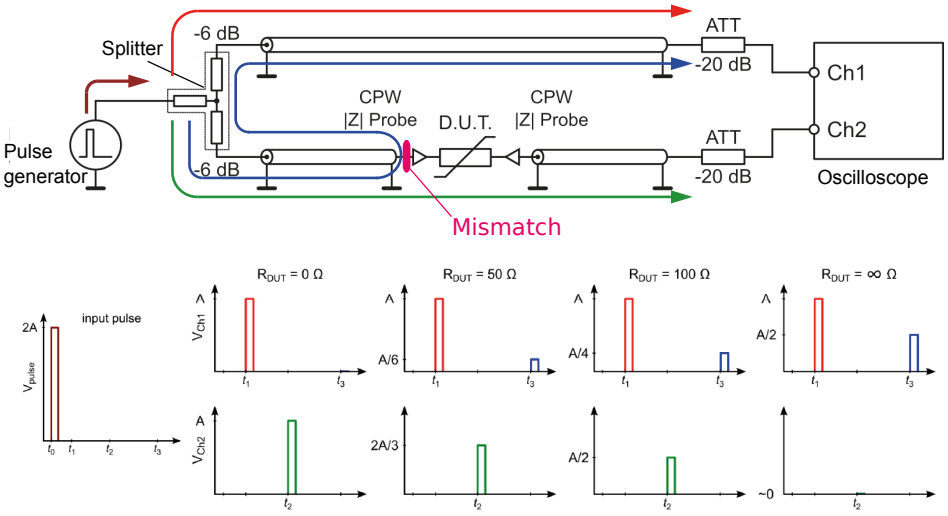


Fig. 20: *(a) Scheme of the setup with pulse generator and signal paths, see text, and (b) simulation of signal transformation at the reference (Ch1) and the DUT (Ch2) output ports in dependence of different resistances* $0\ \Omega \leq R_{\text{DUT}} \leq 100\ M\Omega$*. Delay times caused by the coaxial cable lengths as well as by the 6 dB attenuation by the power splitter are modeled for a real actual setup assembly and result in* $t_0 = 0$ *ns,* $t_1 = 4.5$ *ns,* $t_2 = 17.5$ *ns, and* $t_3 = 19.5$ *ns without taking into account the attenuators (ATT).*

The utilized measurement equipment depends on the frequency regime, e.g. [38]. For the pulses between 100 ns and 0.8 ns a PSPL 2600C Turbo pulse generator deliver 0 dB attenuation a bipolar output amplitude of $-45\ \text{V} \leq \hat{V} \leq 50$ V with pulse rise/fall times $t_{\text{rise}} \approx 250$ ps and $t_{\text{fall}} < 800$ ps. The output attenuation is adjustable from 0 to 70 dB in discrete 1 dB steps. The signal from the generator output is divided via a PSPL 5333 power splitter which introduces a 6 dB attenuation per port. The oscilloscope Tektronix DPO73304D, 4 analog channels, 33 GHz, 100 GS/s single shot, provides sufficient bandwidth and sampling rate for the application. In order to avoid any damage of the oscilloscope input circuitry fixed 20 dB attenuators are inserted. For shorter pulses between 1 ns and 100 ps slight modifications in terms of pulse generation and signal routing are required. A PSPL 12050 12.5 GHz pattern generator produces pulses with a length down to 76 ps and 25 ps rise/fall time. However, a single pulse must be isolated from the continuous pattern and amplified from the given output amplitude in the range between 0.25 V_{PP} and 2 V_{PP} to a level which is sufficient for the ReRAM switching operation. For this reason it makes sense to give up the power splitter, i.e. an additional 6 dB damping. and analyze only the signal passing through the DUT. The reference is measured directly at the trigger module output, prior to coupling to the DUT.

5.3 Design of integrated resistive switching cell

In HF systems signals are transmitted in different ways, e.g. waveguides, two wire lines, coaxial lines, or planar transmission lines. Planar transmission lines are available in form of striplines, microstrip lines, slotlines, coplanar waveguides or further types with related geometries. The preferential design for the ReRAM prototyping is the coplanar waveguide (CPW) [40] enabling a compact device integration, easy feasibility and low costs. The concept is generally usable from DC to 60 GHz. Unlike microstrip and other waveguide types, the CPW places the signal and ground on the same layer of the substrate, see Fig. 21. The CPW approach is CMOS compatible for device integration [36]. Hence the CPW waveguide metals can be deposited and structured on a silicon wafer by photolithography together with active materials exhibiting resistive switching, forming an HF compatible prototype of a planar ("sufficient thin") ReRAM cell. Parameters of CPW are set by the physical properties and the geometry of the used substrates and electrode materials. The impedance $Z_0 = 50\ \Omega$ is determined by the substrate thickness h, the permittivity ε_r, the metal thickness t, the stripe width w, and the gap width s, e. g. [39], see Fig. 21a,b. The tapered shape of the waveguide guarantees a constant impedance when the signal is guided from the large pads for coaxial probe tips to the small dimensions of the active device in the middle conductor (Fig. 21c).

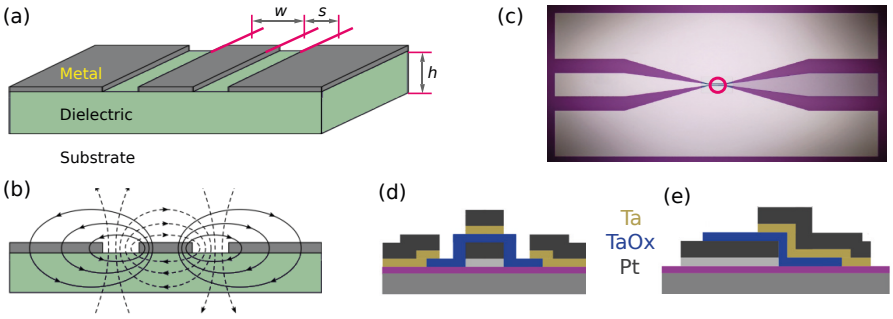


Fig. 21: *(a) A coplanar waveguide structure: a single layer metal pattern on a dielectric substrate, comprising middle signal stripe and two ground planes. (b) Distribution of electric and magnetic field in CPW adapted from [36], (c) tapered line ($w = 100\,\mu m$, $s = 50\,\mu m$) with an active ReRAM element inside the red circle ($w = 5\,\mu m$, $s = 2.5\,\mu m$), (e),(f) side-views of a Pt-TaO$_x$-Ta ReRAM cell with an area of overlapping electrodes of $A = w \times 10\,\mu m$.*

The measurement setup for the ReRAM cell in series with the tapered CPW is represented by the equivalent circuit as shown in Fig. 22. The resistive switching element is described by a resistor and capacitor in parallel, additional resistors R_{TE} and R_{BE} as well as capacitors C_{i} result from parasitic effects, e. g. by electrodes and by the current through the ReRAM device capacitance associated with the geometry and material stack of the memory cell. The current through the resistive switching cell I_{DUT} and the voltage drop across the cell V_{DUT} can be obtained from the single-shot measurement of V_{inc} and V_{trans}.

$$V_{\mathrm{inc}} = V_{\mathrm{ref}} + V_{\mathrm{trans}} \qquad V_{\mathrm{DUT}} = 2\ (V_{\mathrm{inc}} - V_{\mathrm{trans}}) \qquad I_{\mathrm{DUT}} = \frac{V_{\mathrm{trans}}}{Z_0}. \tag{8}$$

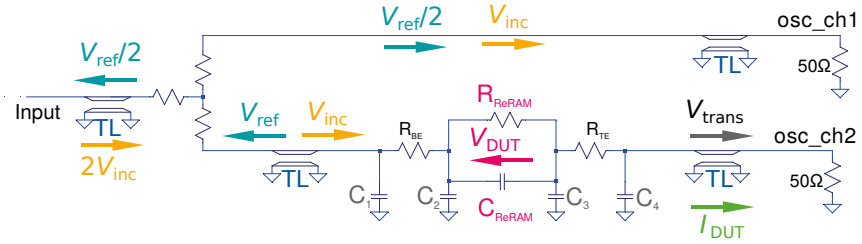


Fig. 22: *Equivilant circuit of a ReRAM device including signal propagation.*

5.4 Ultra-fast resistive switching of tantalum oxide cell

Exemplarily, results of picosecond and nanosecond SET pulse measurements are shown for a CPW TaO_x resistive switching device [38]. The scheme of Fig. 23 illustrates the procedure. In order to fix a defined starting condition the cell was switched to the OFF state by a default RESET sweep. The SET pulse was varied with respect to amplitude and length constant. After each SET process the device was read out with a 100 mV bias voltage and subsequently a complete SET-RESET sweep.

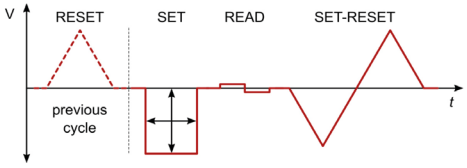


Fig. 23: *Time signal scheme of the SET switching procedure [38].*

Fig.24 a shows a series of current transients of the output signal acquired on the 50 Ω oscilloscope's input when 10 ns pulses of different levels are applied on a 10 × 20 μm^2 TaO_x CPW ReRAM device. By subsequent voltage sweeps $R(V)$ as depicted in Fig.24 b the resistance state is analyzed in which the initial point of the curves (-0.05 V) denotes the resistance level immediately after the specific applied pulse. The actual resistive switching event t_{SET} is matched

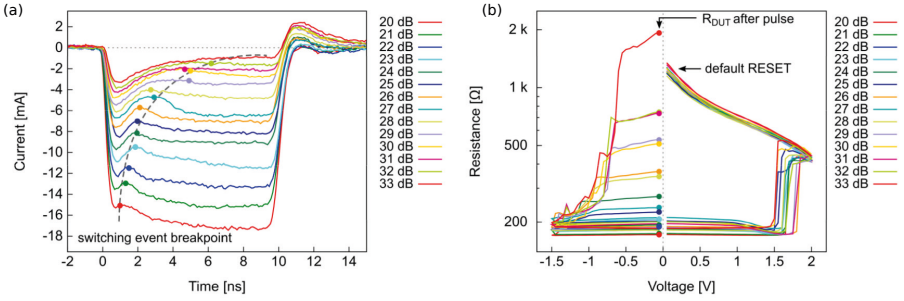


Fig. 24: *SET pulses with different amplitudes. The levels in dB denote the attenuation of -45 V (0 dB) signal magnitude, and (b) plots of the subsequent I-V sweeps following the procedure of Fig. 23 [38].*

to the breakpoint in each transient current response marked by points. t_{SET} is decreasing with increasing the voltage level. The signal overshoots at the beginning and the end of the pulse are caused by the device capacitance, i. e. the electrode overlap in the CPW tapered area. When the pulse amplitudes do not exceed a switching threshold voltage, the cell stays in the OFF state and the passing signal is affected only by the cell capacitance. For a 10 ns pulse length the threshold value is -1 V (att. ≥33 dB). At pulse amplitudes $-1.1\ \mathrm{V} \leq V_{\mathrm{pulse}} \leq -2.9\ \mathrm{V}$, the cell is switched to an intermediate resistive state, with resistance value proportional to the signal attenuation level. The feasibility to tune the ON resistance value with the means of the signal amplitude was already introduced as multi-level switching. For $V_{\mathrm{pulse}} \geq -3.2$ V (att. ≤23 dB) the cell is set to a "hard ON" state.

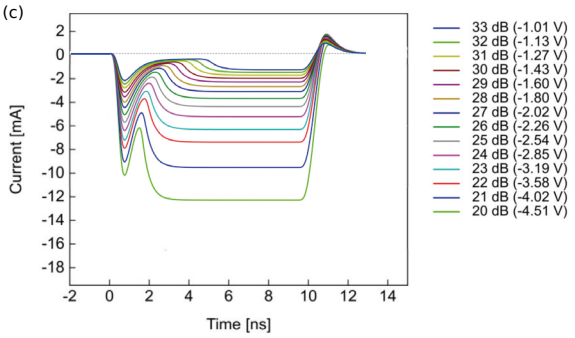


Fig. 25: *Results of spice transient analysis. The variable pulse amplitude determines the switching delay, switching speed and R_{DUT} after the pulse, see also Fig. 24. The resistance change is controlled by a linear changing time variable. The depicted transients show current over a 50 Ω termination resistor [38].*

A spice model (LTspice [41]) is used to simulate the HF response of a ReRAM cell integrated in a tapered CPW line [38]. The input pulses are realized by rectangular pulses combined with phase shifted sine half periods as rise/fall edges similar to the actual generator signal form. The TL are considered as lossless with a time delay 4.5 ns/m. The ReRAM device is modelled by the actual resistance R, the capacitance of the electrode overlapping area C_{ReRAM}, the series resistance of the electrodes $R_{\mathrm{BE,TE}}$, and by the distributed stray capacitance C_{i} as shown in Fig. 22. Fig. 25 illustrates the simulation results of 10 ns pulse series which are in good agreement to the experimental data of Fig. 24.

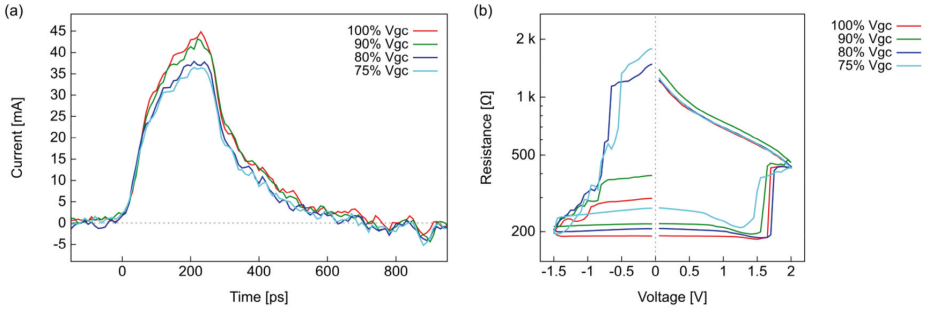


Fig. 26: *(a) Measurements of 250 ps SET pulses on a 10 × 20 μm^2 TaOx CPW ReRAM device with different amplitudes, and (b) plots of the subsequent voltage sweeps following the procedure of Fig. 23 [38].*

Resistive switching also present at pulse widths down to the order of 10^2 ps, with multi-level characteristics even [38]. Unfortunately, the switching event is not observable directly due to the fact that the switching event occurs already during the pulse rise time (dominating effect of the cell capacitance). A readout is needed to confirm a resistance change after a sub-ns pulse was applied at the cell. The results are plotted in Fig. 26.

6 Conclusion

In this contribution electrical characterization of resistive switching memory cells with focus on the device behavior upon current compliance, and device time dependence down to picosecond timescale is addressed. For RRAM devices, there is a need for proper dynamical device characterization which are beyond simplistic analysis available by the means of commercial tools. Impedance matching, prevention of parasitics, and detailed know-how of driver performance is crucial for fixing device parameters.

The performance of the presented different measurement methods is summarized in a common plot of Fig. 27. The results show the switching (SET) time as a function of voltage pulse amplitude for a single TaO_x cell over a timescale of 15 decades. This demonstrates on the one hand the potential of the electrical characterization setup, and on the other hand, the excellent switching kinetics of TaO_x memory cells which are able to overcome the voltage-time dilemma of RRAM devices.

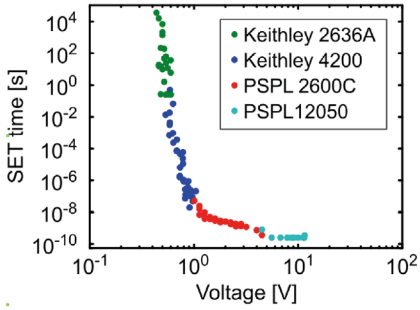


Fig. 27: *SET switching time of 15×20 μm^2 TaO_x CPW resistive switching device as a function of voltage. The measurements were conducted on a same device using multiple measurement setups in order to cover broad timescale. The switching kinetics extends over 15 decades at the timescale [38].*

References

[1] R. Waser (Ed.), Nanoelectronics and Information Technology Advanced Electronic Materials and Novel Devices, 3rd Edition, Wiley VCH 2016.

[2] L.O. Chua, S.M. Kang, Proc. IEEE 64, 209 (1976).

[3] W.-Y. Chang, J.-H. Liao, Y.-S. Lo, and T.-B. Wu, Appl. Phys. Lett. 94, 172107 (2009).

[4] A. Sawa, T. Fujii, M. Kawasaki, and Y. Tokura, Appl. Phys. Lett. 85, 4073 (2004).

[5] S. Menzel, M. Waters, A. Marchewka, U. Böttger, R. Dittmann, and R. Waser, Adv. Funct. Mat. 21, 4487 (2011).

[6] I. Valov1, R. Waser, J. R. Jameson, and M. N. Kozicki, Nanotechnology 22, 254003 (2011).

[7] C. Schindler, G. Staikov, R. Waser, Appl. Phys. Lett. 94, 072109 (2009).

[8] D. L. Losee, J. Appl. Phys. 46, 2204 (1975).

[9] T. Menke, P. Meuffels, R. Dittmann, K. Szot, and R. Waser, J. Appl. Phys. 105, 066104 (2009).

[10] D. Ielmini and R. Waser (eds.), Resistive Switching: From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications, Wiley VCH 2016.

[11] Keithley, Technical Information: Source Measurement Unit (SMU) Instruments (2015).

[12] D. Ielmini, F. Nardi, and S. Balatti, IEEE Trans. Electron Devices 59 2049 (2012).

[13] R. Waser, R. Dittmann, G. Staikov, and K. Szot, Adv. Mater. 21, 2632 (2009).

[14] F. Miao, W. Yi, I. Goldfarb, J. J. Yang, M.-X. Zhang, M. D. Pickett, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, ACS Nano 6, 2312 (2012).

[15] S. Schmelzer, Ultra Thin Oxide Films for Dielectric and Resistive Memory Applications, PhD thesis RWTH Aachen (2013).

[16] M. J. Lee, C. B. Lee, D. Lee, S. R. Lee, M. Chang, J. H. Hur, Y. B. Kim, C. J. Kim, D. H. Seo, S. Seo, U. I. Chung, I. K. Yoo, and K. Kim, Nat. Mater. 10, 625 (2011).

[17] K. M. Kim, B. J. Choi, M. H. Lee, G. H. Kim, S. J. Song, J. Y. Seok, J. H. Yoon, S. Han, and C. S. Hwang, Nanotechnology 22, 254010 (2011).

[18] J. Liang, and H.-S. Philip Wong, IEEE Trans. Electron Dev. 57, 2531 (2010).

[19] J. P. Strachan, A. C. Torrezan, F. Miao, M. D. Pickett, J. J. Yang, W. Yi, G. Medeiros-Ribeiro, and R. S. Williams, IEEE Trans. Electron Dev. 60, 2194 (2013).

[20] R. Symanczyk, R. Dittrich, J. Keller, M. Kund, G. Mueller, B. Ruf, P.-H. Albarede, S. Bournat, L. Bouteille, A. Duch, Proceedings of NVMTS None, 71 (2007).

[21] Y. M. Lu, M. Noman, W. Chen, P. A. Salvador, J. A. Bain, and M. Skowronski, J. Phys. D Appl. Phys. 45, 395101 (2012).

[22] K. Kinoshita, K. Tsunoda, Y. Sato, H. Noshiro, S. Yagaki, M. Aoki, and Y. Sugiyama, Appl. Phys. Lett. 93, 033506 (2008).

[23] Keithley Instruments Inc., Series 2600B System SourceMeter Instrument Reference Manual (2012).

[24] K. Fleck,U. Böttger, R. Waser, S. Menzel, IEEE Electron Device Lett. 35 (2014) 924 - 926.

[25] S. Yu, Y. Wu, and H.-S. Philip Wong, Appl. Phys. Lett., 98, 103514 (2011).

[26] J. P. Strachan, A. C. Torrezan, G. Medeiros-Ribeiro, R. S. Williams, Nanotechnology 22, 505402 (2011).

[27] D. Ielmini, F. Nardi, and S. Balatti, IEEE Trans. Electron Devices 59 2049 (2012).

[28] G. Bruns, Electronic switching in phase-change materials, Ph.D. thesis, RWTH Aachen (2012).

[29] C. Hermes, M. Wimmer, S. Menzel, K. Fleck, G. Bruns, M. Salinga, U. Bttger, R. Bruchhaus, T. Schmitz-Kempen, M. Wuttig, and R. Waser, IEEE Electron Dev. Lett. 32, 1116 (2011).

[30] G. Bruns, P. Merkelbach, C. Schlockermann, M. Salinga, M. Wuttig, T. D. Happ, J. B. Philipp, and M. Kund, Appl. Phys. Lett.95, 043108 (2009).

[31] Analog Devices, Dealing with High-Speed Logic, Tutorial MT-097 (2009).

[32] L. Frenzel, Back to Basics: Impedance Matching (2011).

[33] D. M. Pozar, Microwave Engineering, John Wiley & Sons (2012).

[34] Hewlett-Packard, S-Parameter Techniques, Application Note 95-1 (1997).

[35] Agilent Technologies, Understanding the Fundamental Principles of Vector Network Analysis, Application Note (2012).

[36] I. Rosu, Microstrip, Stripline, and CPW Design (2014).

[37] A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, Nanotechnology, vol. 22, 485203, 2011.

[38] V. Havel, PhD thesis, RWTH Aachen (exp. 2016).

[39] E. Chen and S. Y. Chou, IEEE Trans. Microw. Theory Tech. 45, 939 (1997).

[40] C. P. Wen, IEEE Trans. Microw. Theory Tech. 17, 1087-1090 (1969).

[41] Linear Technology Corporation, http://www.linear.com.

C 2 X-Ray Diffraction and Scattering

Uwe Klemradt
RWTH Aachen University, Germany

Contents

1 Introduction **2**

2 X-ray Diffraction from Bulk Crystals Revisited **3**
2.1 Elementary Scattering Theory 3
2.2 Laue and Bragg Equations 4

3 Small Angle X-ray Scattering **9**

4 X-ray Optics under Grazing Incidence **11**

5 Grazing Incidence Small Angle X-ray Scattering **12**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

A deeper understanding of the interrelation between electrically induced resistance changes and modifications of the atomic and electronic structure are prerequisite for the development of valence change memories with predictable memristive device performance and stability [1]. At first glance, it may appear surprising that X-rays can be used to contribute to the physics of defect-related, localized transport phenomena in thin films. However, X-rays can be used for much more than determining the average lattice structure of bulk crystals. Under grazing incidence, all X-ray techniques become surface sensitive [2], and the X-ray penetration depth of a few nm under total reflection is well suited to the nanoscale films employed in devices. Moreover, the size and shape of non-periodically spaced filamentary structures can be investigated by small angle X-ray scattering, a technique widely used in soft matter research [3]. Last not least, the advent of 3rd generation synchrotron sources has brought about the possibility to focus X-ray beams down to the nanoscale, thus allowing also extreme spatial resolution with scattering methods [4].

The outline of this lecture is as follows. First, we start with a short repetitorium of conventional X-ray scattering from crystals to lay a foundation for concepts like cross section and charge scattering before deriving well-known results like the Laue and Bragg equations. In the following chapter, small angle X-ray scattering (SAXS) is introduced as a standard technique for the structural characterization of objects randomly distributed in a matrix, for example colloids in solution. Emphasis is laid on the proper generalization of results from - and analogies with - X-ray diffraction from crystals. The fourth section sums up the principles of X-ray optics under grazing incidence, which leads in a natural way to near-surface sensitivity by total external reflection. The last section introduces, finally, grazing incidence small angle X-ray scattering as a tool combining grazing incidence with SAXS for the study of inhomogeneities in thin films. The section concludes with an application to filament formation in memristive devices.

In the world of nanostructured materials, diffraction and scattering phenomena caused by the interaction of radiation with matter provide useful probes for the investigation of structural properties at atomistic and mesoscopic length scales. Although the interaction processes are usually understood as diffraction of waves or collision of particles, quantum physics tells us that the classical distinction between particles and waves is meaningless, since quantum objects like photons, electrons, or neutrons behave according to the laws of quantum physics with particle *and* wave like properties. Only in certain situations they can be described approximately as classical particles *or* waves [5]. In this sense, the words particle and wave must be used exchangeably for the description of what is a generic quantum process. This usage of words must be kept in mind also for the present lecture, when the classical wave picture is predominantly, but not exclusively used to describe scattering processes.

2 X-Ray Diffraction from Bulk Crystals Revisited

2.1 Elementary Scattering Theory

Scattering experiments usually consist of a beam of incoming particles (photons, electrons, neutrons,…), which are deflected at a target and then registered at a detector, whose position is defined by its direction relative to the incident beam, described by the two angles (ϑ, φ), which we will denote by the shorthand Ω (solid angle). For a parallel and monoenergetic beam of incident particles, the incoming wavefunction can be represented as a plane wave

$$\psi_{inc}(\boldsymbol{r}) = A\, e^{i\boldsymbol{kr}}, \tag{1}$$

where A is a normalization factor and **k** the wavevector, which is proportional to the momentum $\hbar\boldsymbol{k}$ of the particles. If the interaction is elastic and has a short range, the scattered wave outside the interaction region of the target (located at $\boldsymbol{r} = 0$) can be described by a spherical wave

$$\psi_{scat}(\boldsymbol{r}) = B(\Omega)\frac{e^{ikr}}{r}. \tag{2}$$

(cf. Fig. 1). Elastic interactions do not change the energy of the incoming particle and the absolute value of the wavevector: $k' = k = |\boldsymbol{k}| = 2\pi/\lambda$. The factor $B(\Omega)$ contains the information about the scattering process [6]. The total wavefunction outside the interaction region can be written as

$$\psi_{total}(\boldsymbol{r}) = \psi_{inc}(\boldsymbol{r}) + \psi_{scat}(\boldsymbol{r}) = A\left(e^{i\boldsymbol{kr}} + f(\Omega)\frac{e^{ikr}}{r}\right), \tag{3}$$

where $f(\Omega) = B(\Omega)/A$ has the dimension of a length and is often known as scattering amplitude or "scattering length", a concept which will be discussed in more detail below.

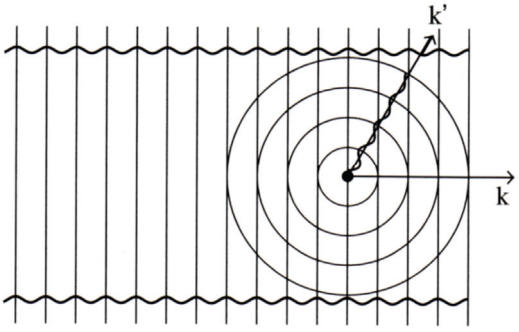


Fig. 1: *Schematic view of the scattering process from a fixed target. The thin curves indicate lines of equal phase and equal amplitude and the thick wavy curves indicate the amplitudes of the incident and scattered wave that have the wavevectors k and k' [6].*

An important quantity for the experimental discussion of interaction processes is the concept of a cross section, which is defined as the ratio of the *current* of scattered particles and the *current density* of incident particles. A schematic view of the situation is given in Fig. 2. The interpretation of $|\psi_{inc}(\boldsymbol{r})|^2 = |A|^2$ and $|\psi_{scat}(\boldsymbol{r})|^2 = |B(\Omega)|^2/r^2$ as (probability) densities of the incident and scattered particles and the volumes defined in Fig. 2 can be used to obtain these currents. The number of incident particles which cross the surface ΔS during time Δt is

given by $N = v\,\Delta t\,\Delta S\,|A|^2$, and the number of scattered particles which reach the detector through an opening angle $\Delta\Omega$ at a distance r from the target during time Δt is given by $\Delta N = v\,\Delta t\,r^2\Delta\Omega\,|B(\Omega)|^2/r^2$. The ratio of the current of scattered particles and the current density of incident particles is

$$\Delta\sigma = \frac{\Delta N}{N/\Delta S} = |B(\Omega)|^2\ [A]^2\,\Delta\Omega, \tag{4}$$

from which

$$\sigma = \int d\Omega\,|f(\Omega)|^2 \tag{5}$$

follows. This quantity has the dimension of an area and is called cross section of the scattering, whereas

$$\frac{d\sigma}{d\Omega} = |f(\Omega)|^2 \tag{6}$$

is called the differential cross section. Obviously, counting the number of particles at the detector only yields information on the scattering intensity $|f(\Omega)|^2$, but not on the scattering amplitude f(Ω). This is known as the phase problem that exists in the interpretation of scattering experiments, since the information on the phases is lost.

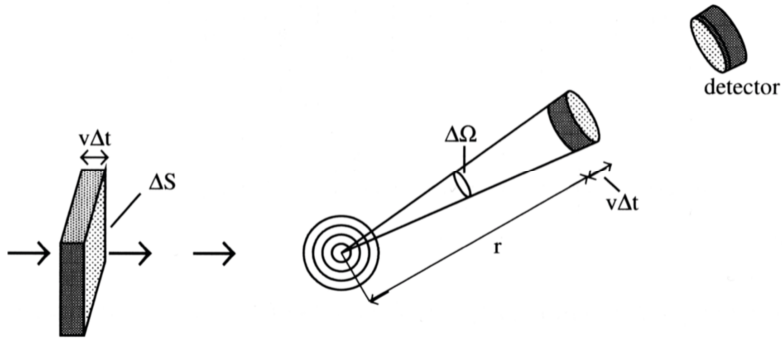


Fig. 2: *The number N of incident particles, which cross the surface ΔS during time Δt, are contained in a parallelepiped of base ΔS and height $v\,\Delta t$, where v stands for the particle velocity. The number ΔN of particles, which reach the detector through an opening angle $\Delta\Omega$ at distance r from the target during time Δt are contained in a conic volume of base $r^2\,\Delta\Omega$ and height $v\,\Delta t$. Adapted from [6].*

2.2 Laue and Bragg Equations

The scattering of X-rays occurs from the total electron density ρ of a sample. The amplitude $f(\Omega)$ of elastic scattering (Thomson scattering) can be written as a phase-adjusted sum that depends only on the scattering vector $\boldsymbol{Q} := \boldsymbol{k}' - \boldsymbol{k}$:

$$f(\boldsymbol{Q}) = -r_0 \int d^3r \; e^{-i\boldsymbol{Q}\boldsymbol{r}} \sum_{n=1}^{N} \rho_n(\boldsymbol{r} + R_n), \tag{7}$$

where $r_0 = 2{,}818 \cdot 10^{-15}$ m is the classical electron radius, the minus sign signifies a phase shift of 180°, and the scattering density is broken down into a sum of atomic densities. R_n denotes the atomic positions of a set of N atoms. By a shift of the integration variable, the last equation is changed into

$$f(\boldsymbol{Q}) = -r_0 \sum_{n=1}^{N} e^{i\boldsymbol{Q}\boldsymbol{R_n}} \int d^3r \; e^{-i\boldsymbol{Q}\boldsymbol{r}} \rho_n(\boldsymbol{r}) = -r_0 \sum_{n=1}^{N} e^{i\boldsymbol{Q}\boldsymbol{R_n}} \; f_n(\boldsymbol{Q}). \tag{8}$$

This shows that the scattering amplitude can be written as a sum over phase factors, which depend on the atomic positions, multiplied by single atom scattering amplitudes. The $f_n(\boldsymbol{Q})$ are also called atomic form factors, since the detailed shape of the electron distribution determines the interference at finite scattering angles - in other words, $f_n(\boldsymbol{Q})$ stands for the angular-dependent 'scattering strength' of the atom at position *n*. If all atomic form factors are identical, the scattering intensity is proportional to $N\, S(\boldsymbol{Q})\, |f(\boldsymbol{Q})|^2$, where the structure factor $S(\boldsymbol{Q})$ is defined as

$$S(\boldsymbol{Q}) = \frac{1}{N} \left| \sum_{n=1}^{N} e^{i\boldsymbol{Q}\boldsymbol{R_n}} \right|^2 = \frac{1}{N} \sum_{n,n\prime} e^{i\boldsymbol{Q}(\boldsymbol{R_n} - \boldsymbol{R}_{n\prime})}. \tag{9}$$

The structure factor describes the interference effects of the scattering at the atomic positions $\boldsymbol{R}_n$ and $\boldsymbol{R}_{n\prime}$.

For crystalline solids the consideration of the structure factor is particularly relevant. In such systems the atoms are arranged on the points of a lattice. For a Bravais lattice, which is characterized by a primitive unit cell containing only one atom, the lattice vectors $\boldsymbol{R}_n$ can be represented as combinations of three linearly independent basis vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$ in the form $\boldsymbol{R}_n = n_1\boldsymbol{a_1} + n_2\boldsymbol{a_2} + n_3\boldsymbol{a_3}$ with integer numbers n_1, n_2 and n_3.

The situation for the body centered cubic (bcc) and face centered cubic (fcc) lattices, which are important structures for many elemental metals, is illustrated in Fig. 3. In Bravais lattices the sum over the phase factors can be written as

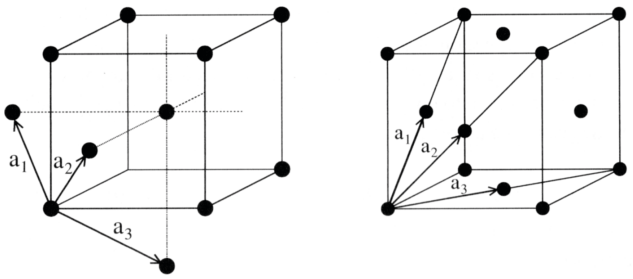


Fig. 3: *Body-centered cubic (left) and face-centered cubic lattices with a choice of possible basic vectors [7]. The conventional (non-primitive) cubic unit cell contains two and four atoms, respectively.*

$$\sum_{n} e^{i\boldsymbol{Q}\boldsymbol{R_n}} = \sum_{n_1=1}^{N_1} e^{in_1\boldsymbol{Q}\boldsymbol{a_1}} \left(\sum_{n_2=1}^{N_2} e^{in_2\boldsymbol{Q}\boldsymbol{a_2}} \left(\sum_{n_3=1}^{N_3} e^{in_3\boldsymbol{Q}\boldsymbol{a_3}} \right) \right) \tag{10}$$

with $N = N_1\, N_2\, N_3$ being the total number of atoms. By summing up the geometric series

$$\sum_{n_1=1}^{N_1} e^{in_1\boldsymbol{Q}\boldsymbol{a_1}} = \frac{\sin(N_1\boldsymbol{Q}\boldsymbol{a_1}/2)}{\sin(\boldsymbol{Q}\boldsymbol{a_1}/2)}\, e^{i((N_1+1)\boldsymbol{Q}\boldsymbol{a_1}/2)} \tag{11}$$

the structure factor $S(\boldsymbol{Q})$ is then obtained as

$$S(\boldsymbol{Q}) = \frac{1}{N}\left|\sum_{n=1}^{N} e^{i\boldsymbol{Q}\boldsymbol{R_n}}\right|^2 = \frac{1}{N}\frac{\sin^2\left(N_1\frac{\boldsymbol{Q}\boldsymbol{a_1}}{2}\right)}{\sin^2\left(\frac{\boldsymbol{Q}\boldsymbol{a_1}}{2}\right)}\frac{\sin^2\left(N_2\frac{\boldsymbol{Q}\boldsymbol{a_2}}{2}\right)}{\sin^2\left(\frac{\boldsymbol{Q}\boldsymbol{a_2}}{2}\right)}\frac{\sin^2\left(N_3\frac{\boldsymbol{Q}\boldsymbol{a_3}}{2}\right)}{\sin^2\left(\frac{\boldsymbol{Q}\boldsymbol{a_3}}{2}\right)} \tag{12}$$

From the maxima of the curves shown in Fig. 4, it can be deduced that large contributions to $S(\boldsymbol{Q})$ arise only from values of $\boldsymbol{Q}$ which are determined by the conditions

$$\boldsymbol{Q}\boldsymbol{a_1} = 2\pi\, n_1, \qquad \boldsymbol{Q}\boldsymbol{a_2} = 2\pi\, n_2, \qquad \boldsymbol{Q}\boldsymbol{a_3} = 2\pi\, n_3 \tag{13}$$

with arbitrary integer values $\boldsymbol{n_1}, \boldsymbol{n_2}, \boldsymbol{n_3}$. These conditions can be satisfied by all Q-vectors

$$\boldsymbol{Q} = h_1\boldsymbol{b_1} + h_2\boldsymbol{b_2} + h_3\boldsymbol{b_3} \tag{14}$$

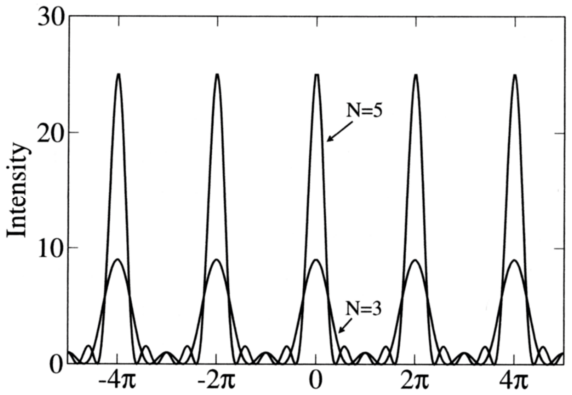


Fig. 4: *The function* $sin^2(Nx/2)\,/\,sin^2(x/2)$ *for* $N = 3$ *and* $N = 5$. *The maximum intensity is given by* N^2 *and the peak width is approximately given by* $2\pi/N$. *For large N, the peaks approach* δ*functions, and the side maxima disappear. [6]*

with integer numbers $\boldsymbol{h_1}, \boldsymbol{h_2}$ and $\boldsymbol{h_3}$, if the vectors $\boldsymbol{b_1}, \boldsymbol{b_2}$ and $\boldsymbol{b_3}$ obey the condition $\boldsymbol{a_i}\,\boldsymbol{b_j} = \boldsymbol{2\pi}\,\boldsymbol{\delta_{ij}}$. The vectors $\boldsymbol{b_1}, \boldsymbol{b_2}$ and $\boldsymbol{b_3}$ can be explicitly constructed by

$$\boldsymbol{b_1} = \frac{2\pi}{V}(\boldsymbol{a_2} \times \boldsymbol{a_3}),\ \ \boldsymbol{b_2} = \frac{2\pi}{V}(\boldsymbol{a_3} \times \boldsymbol{a_1}),\ \ \boldsymbol{b_3} = \frac{2\pi}{V}(\boldsymbol{a_1} \times \boldsymbol{a_2}), \tag{15}$$

where $V = \boldsymbol{a_1} \cdot (\boldsymbol{a_2} \times \boldsymbol{a_3})$ is the volume of the unit cell. They are the basis vectors for the reciprocal lattice, which is associated with the real space lattice:

$$\boldsymbol{G}_{hkl} = h\boldsymbol{b_1} + k\boldsymbol{b_2} + l\boldsymbol{b_3}, \tag{16}$$

with integer values h, k, l. For very large periodic assemblies of atoms (e.g., macroscopic crystals), the ratio of the sin functions in Eq. (12) can be replaced by a δ function in the limit $N \to \infty$, yielding a structure factor of the form

$$S(\boldsymbol{Q}) = \sum_{h,k,l} \delta(\boldsymbol{Q} - \boldsymbol{G}_{hkl}). \tag{17}$$

This shows that scattering occurs only if the scattering vector $\boldsymbol{Q}$ coincides with a reciprocal lattice vector $\boldsymbol{G}_{hkl}$, hence the importance of the reciprocal lattice for waves interacting with crystals. Since real crystals are neither infinite nor consist of exactly periodic repetitions of the unit cell, not only sharp Bragg peaks at $\boldsymbol{Q} = \boldsymbol{G}_{hkl}$ are observed, but also diffuse scattering exists between the peaks, which contains information on the real structure of the crystals.

The main peaks are of course the well-known Bragg reflections. Their positions in reciprocal space reveal the metrics of the unit cell (lattice constants a, b and c and unit cell angles α, β and γ). The experimental width of the Bragg peaks is not only determined by the size of the coherently scattering volume (mosaic distribution, internal strain, etc.), but depends also on the experimental resolution.

The visualization of the scattering condition Eq. (17) leads to the famous Ewald construction. For simplicity, we will take a simple cubic lattice as an example. Fig. 5 shows a part of a plane perpendicular to the [001] direction, resulting in a quadratic arrangement of reciprocal lattice dots. According to Eq. (17), they represent possible Bragg reflections if the scattering vector $\boldsymbol{Q}$ points to a point of the reciprocal lattice. We now use the notation $\boldsymbol{k}_i$ for $\boldsymbol{k}$ and $\boldsymbol{k}_f$ for $\boldsymbol{k}'$, yielding $|\boldsymbol{k}_f| = |\boldsymbol{k}_i| = k$ for elastic scattering. The scattering vector is defined by the incoming beam (represented by $\boldsymbol{k}_i$) and the position of the detector (represented by $\boldsymbol{k}_f$):

$$\boldsymbol{Q} = \boldsymbol{k}_f - \boldsymbol{k}_i \tag{18}$$

Since the Laue condition

$$\boldsymbol{Q} = \boldsymbol{G}_{hkl} \tag{19}$$

requires that the scattering vector coincides with a point of the reciprocal lattice, Eq. (18) and (19) can be solved graphically in the following way for a given reciprocal lattice (Ewald construction):

- Note that, according to Eq. (19), $\boldsymbol{Q}$ and the reciprocal lattice must share the same origin. In the Ewald construction, this is ensured by placing the *endpoint* of the incoming wavevector $\boldsymbol{k}_i$ at the origin of reciprocal space. In accordance with Eq. (18), then $\boldsymbol{Q}$ will start from the origin.
- Next, all possible $\boldsymbol{k}_f$ are represented by drawing a circle of fixed radius k around the *starting point* of vector $\boldsymbol{k}_i$. In three dimensions, the circle becomes the so-called Ewald sphere**.**
- According to Eq. (18), all experimentally possible scattering vectors for elastic scattering are found on that circle (sphere in 3D). Their magnitudes range from zero (scattered beam parallel to the incident beam, forward scattering) to $2k$ (scattered beam antiparallel to the incident beam, backscattering).
- Eq. (19) implies that a Bragg reflection is only observed if the Ewald sphere cuts through a reciprocal lattice point. This must be arranged for in an actual experiment, either by rotating the reciprocal lattice (e.g., the crystal, cf. Eq. (15)) or by choosing a different scattering vector, or both.

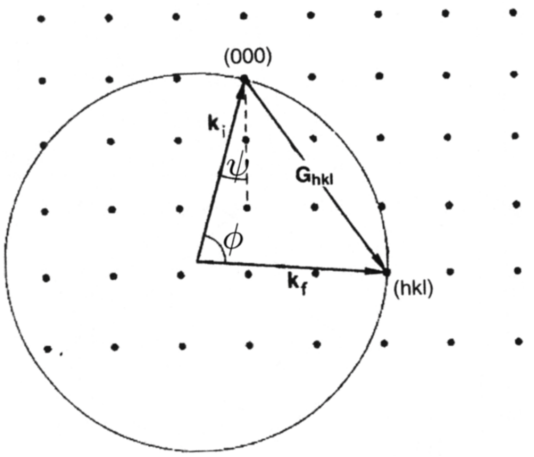


Fig. 5: *Reciprocal space and vector representation for elastic scattering, showing the Ewald construction for Bragg reflection.*

Which scattering vector to choose in order to probe a certain set of crystal lattice planes? Eq. (19) gives the answer by taking the norm, resulting in a scalar version of the vectorial Laue equation, which yields of course the familiar Bragg equation:

$$\begin{aligned} |\boldsymbol{Q}| &= |\boldsymbol{G}_{hkl}| \\ 2k \sin\theta &= 2\pi/d_{hkl} \\ 2d_{hkl} \sin\theta &= \lambda \end{aligned} \tag{20}$$

The Bragg equation (20) states that the detector must be placed an angle of 2θ with respect to the incident beam in order to detect Bragg reflection from lattice planes with a spacing of d_{hkl}. However, this is only a necessary condition; the vectorial nature of Eq. (19) requires a proper alignment of the reciprocal lattice as well. Note that the index hkl related to the (hkl) Bragg reflection is usually not reduced by its greatest common divisor, hence the right-hand side of Eq. (20) does not contain multiples of λ, as often found in elementary derivations of the Bragg condition based on interference from two lattice planes at fixed distance.

3 Small Angle X-ray Scattering

Small angle X-ray scattering (SAXS) is a technique widely used in materials science to characterize and analyse quantitatively agglomerates on a much larger scale than interatomic distances (up to 1000 nm) [3]. This can be achieved with typical X-ray wavelengths of the order of 0.1 nm by exploiting the diffuse halo around the primary beam after interaction with the sample (see Fig. 6). In generalization of Eq. (20), it can easily be understood that the scattering geometry must involve scattering at small angles to resolve very large structures like colloidal particles or polymers: the modulus of $\boldsymbol{Q}$ must be equal to 2π over a characteristic length of the sample under study. Therefore, to study large length scales it is required to make Q very small (hence reciprocal space!), which implies at fixed wavelength small scattering angles. In practice, scattering angles up to about 5° are used together with an area detector.

The incoming monochromatic X-ray beam is scattered elastically from the electrons of the sample; however, only structural inhomogeneities of all kinds (chemical, topological,…) contribute to the diffuse halo, which is detected as a function of $\boldsymbol{Q}$. The primary beam is typically blocked by a beamstop in order to be able to use a highly sensitive 2D detector, which could be damaged by receiving the direct beam. The size of the beamstop, as well as the dimension of the primary beam and the size and distance of the detector, contribute significantly to the experimentally achievable resolution, which ranges for typical experimental setups between 1 and 500 nm.

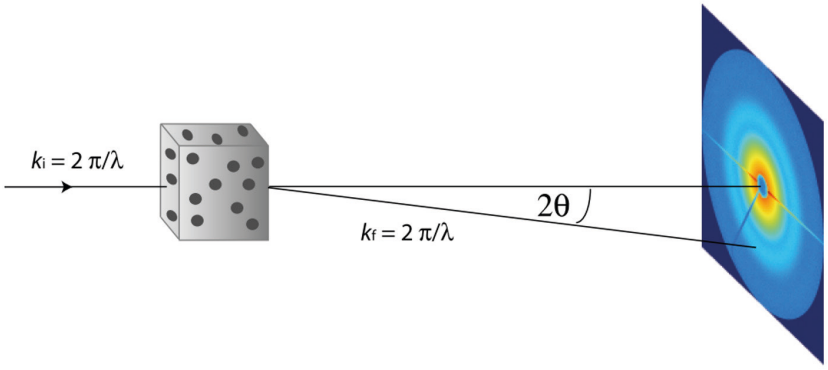


Fig. 6: *The geometry of small angle X-ray scattering.*

In the two phase model, small angle scattering is decribed from (homogeneous) objects embedded in a homogeneous matrix [8]. The scattering contrast is defined as weighted difference of the electron densities

$$\Delta n_f = n_T f_T - n_M f_M \tag{21}$$

with n being the electron density and f a form factor. It can be shown that the detected intensity is proportional to the scattering contrast and the form factor related to the geometry of the scattering objects:

$$\frac{d\sigma}{d\Omega} = \Delta n_f^2 \, V^2 \left| \frac{1}{V} \int_V d^3r \, e^{-i\boldsymbol{Q}\boldsymbol{r}} \right|^2 \tag{22}$$

Isotropic scattering patterns may be caused by isotropic objects, but more frequently they result from anisotropic objects that are randomly oriented in the interaction volume. For isotropic scattering patterns, it is sufficient to integrate the scattered intensity azimuthally at fixed angle 2θ , e.g. for a fixed value of $|\boldsymbol{Q}| = Q$. The typical scattering curve as a function of Q is depicted in Fig. 7, with four distinguishable regimes [9].

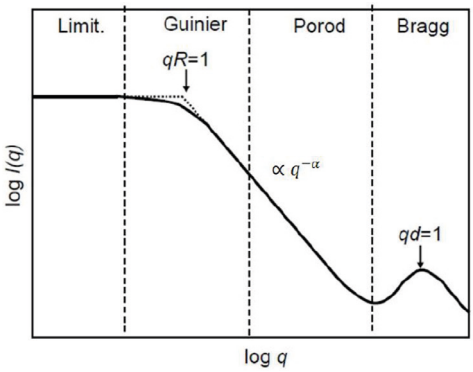


Fig. 7: *Generalized representation of the integrated scattering intensity in small angle scattering with four different regimes.*

The asymptotic part of the scattering curve for $Q \to 0$ is dominated by forward scattering. In the Guinier regime, a transition to falling intensities occurs, characterized by a length scale R which is representative of the maximal size of the scattering objects. The exponent α in the Porod regime depends on the nature of the scattering objects, in particular their geometry. For example, spheres yield a Porod exponent of $\alpha = 4$, whereas plates (discs) yield $\alpha = 3$. However, in many cases pronounced oscillations are observed, and the exponent then relates only to the envelope (cf. Fig. 8). If the objects interact, Eq. (22) must be complemented by a corresponding structure factor.

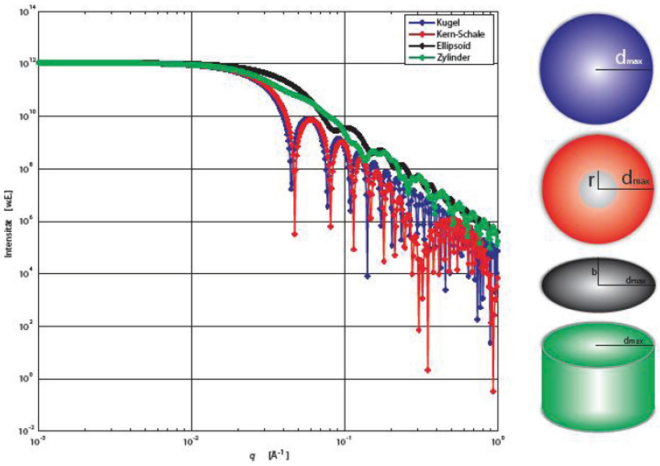


Fig. 8: *Simulated small angle X-ray scattering curves for various particle geometries (blue: spheres, red: spherical core-shell structures, gray: rotational ellipsoids, green: cylinders) [10].*

4 X-ray Optics under Grazing Incidence

The index of refraction is generally dependent on the wavelength of electromagnetic radiation. However, in the hard X-ray regime ($\lambda \approx 1$ Å), the index of refraction is very close to unity with

$$n = 1 - \delta - i\beta. \tag{23}$$

In Eq. (23), the dispersion correction δ is a positive quantity on the order of 10^{-6}, whereas the absorption correction β is for most solid materials on the order of 10^{-5}. Hence, the real part of the index of refraction is below unity, which gives rise to total external reflection, which is of uttermost importance for the analysis of surfaces and near-surface regions as outlined below. It should be noted that the negative sign in front of the dispersion correction δ has physical significance, since it is related to a phase shift of π for forced oscillations above the resonance frequency, which is the case for X-rays interacting with almost all electrons of an atom. By contrast, the sign of the absorption correction β is only determined by the condition that absorption must result in a loss of intensity when waves propagate into a material (e.g., it depends on the orientation of the coordinate system used). It can be shown that $\delta \sim \lambda^2 \rho$, where ρ is the electron density of the material.

According to Snell's law

$$\cos \alpha_i = n \cos \alpha_t, \tag{23}$$

a beam impinging from vacuum or air onto a surface under grazing (shallow) angles α_i is transmitted (refracted) at an angle α_t, as well as reflected under the specular angle $\alpha_r = \alpha_i$ (see Fig. 9). In case of a non-ideal surface, intensity can also be detected under non-specular conditions (see next chapter).

Since $Re(n) < 1$, a Taylor expansion of Snell's law (Eq. (23)) yields the critical angle of total external reflection:

$$\alpha_c = \sqrt{2\delta}. \tag{24}$$

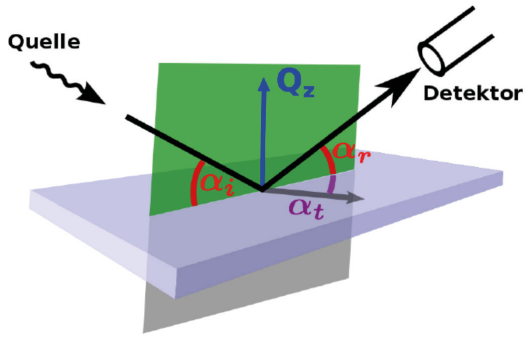


Fig. 9: *Scattering geometry of specular reflectivity. An incident beam directed at a surface under a shallow angle α_i is both transmitted and reflected. Note that the angle α_t is smaller than α_i, allowing for total external reflection.*

For X-rays with $\boldsymbol{\lambda \approx 1}$ Å and common materials, typical values for critical angles are on the order of 0.1°. The existence of a critical angle allows to render all X-ray techniques surface sensitive, since a beam impinging onto a surface below the critical angle is prevented from probing the bulk of the sample. Typical penetration depths are a few nm.

The reflection and transmission of X-rays at interfaces can be described by the Fresnel formulas with the appropriate index of refraction. The reflection and transmission coefficients r and t are defined with respect to the amplitudes A_r of the reflected beam and A_t of the transmitted beam, and normalized with respect to the amplitude A_0 of the incident beam:

$$r = \frac{A_r}{A_0} = \frac{k_{i,z} - k_{t,z}}{k_{i,z} + k_{t,z}} \tag{25}$$

$$t = \frac{A_t}{A_0} = \frac{2k_{i,z}}{k_{i,z} + k_{t,z}}. \tag{26}$$

Note that the reflection and transmission coefficients can be expressed as functions of the z-components (e.g., surface normal) of the k vectors of the incident and transmitted waves. The Fresnel coefficients are not directly experimentally accessible. However, the reflectivity $R = |r|^2$ can be measured fairly easily. Deviations from the ideal Fresnel reflectivity yield information on the surface roughness; in case of thin films interference fringes occur ("Kiessig fringes") that allow a precise determination of the film thickness [11]. The transmittivity $T = |t|^2$ cannot be measured directly. Nevertheless, it shows up in diffuse scattering as pronounced maxima in directions that enclose the critical angle with the surface owing to Eq. (26) exhibiting a peak at the critical angle due to a standing wave field ("Yoneda peak") [12].

5 Grazing Incidence Small Angle X-ray Scattering

Grazing incidence small angle X-ray scattering (GISAXS) combines the sensitivity of small angle X-ray scattering (SAXS) with respect to inhomogeneities with the surface sensitivity of grazing incidence (GI) as outlined in the previous section. All inhomogeneities related to the surface or a thin film probed by the X-rays contribute to an extended (diffuse) non-specular signal (cf. Fig. 10). Surface inhomogeneities include clusters deposited at the surface, but also surface roughness; thin film inhomogeneities are for example precipitates, pores, and interfacial roughness. In principle, scattering contributions from all these inhomogeneities can be calculated using the Distorted Wave Born Approximation (DWBA) [13]. It is difficult to separate the various scattering contributions experimentally if many of them occur simultaneously. However, in practice it is frequently the case that one or two types of inhomogeneities dominate the scattering signal, which then can be successfully evaluated by simulation. The great advantage of GISAXS experiments is that they are non-destructive and hence can be performed in-situ. Moreover, they provide access to buried parts of the sample and give representative results due to averaging over a large area of the sample, as the footprint of the beam is large under grazing incidence.

GISAXS patterns are frequently evaluated by means of cuts in certain directions. It should be noted that the vertical direction indicated q_z in Fig. 10 contains at fixed incident angle α_i not only the q_z-direction, but also (small) contributions of q_x. Sample properties related to the growth direction like the rms roughness, film thickness or the height of clusters lead to characteristic patterns in q_z-direction, whereas lateral properties of the sample like the width of clusters or their correlation length show up in scattering features in the q_y-direction. Similar to diffraction, the measured intensity can be broken down in a form factor $F(\boldsymbol{Q})$ and a structure factor $S(\boldsymbol{Q})$ according to

$$I \sim |F(\boldsymbol{Q})|^2 \, S(\boldsymbol{Q}). \tag{27}$$

Whereas the form factor describes the geometry of the inhomogeneities in analogy to Eq. (22), the structure factor describes the tendency for separation or clustering according to the specific interaction.

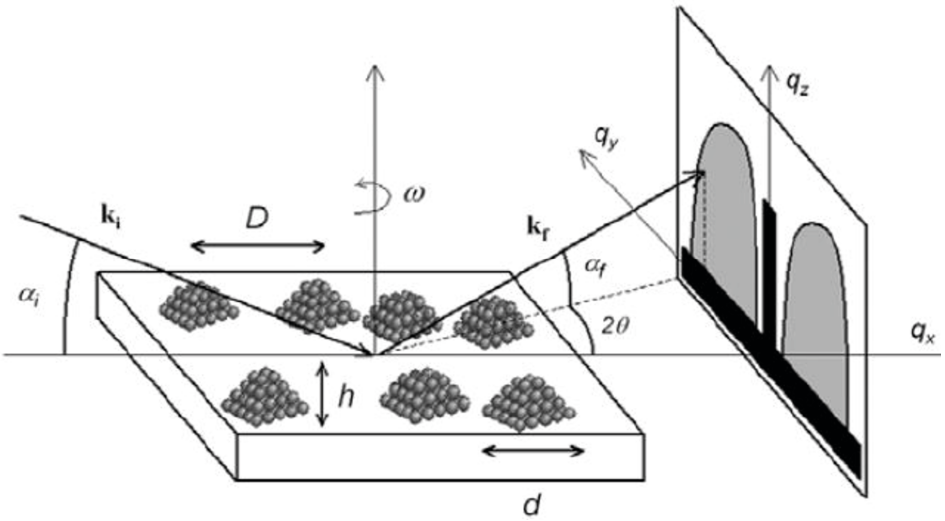


Fig. 10: *Schematics of a GISAXS experiment [13].*

Form factors can be calculated within the DWBA, which is essentially a perturbation theory for X-ray scattering from thin films and interfaces. Whereas in the conventional Born approximation (BA) the perturbation corrections are calculated with plane waves (the analytical solution for wave propagation in free space), the DWBA uses the Fresnel wave fields (the analytical solution for wave propagation in the presence of ideal surfaces). Hence, the DWBA is much more suited than the BA to calculate perturbation corrections to the reflection from ideal interfaces, and nowadays it is firmly established for the theoretical description of GISAXS patterns. The DWBA effective form factor $\tilde{F}$ is based on four contributions, which are depicted in Fig. 11:

$$\tilde{F}(\boldsymbol{Q}) = \tilde{F}\left(q_y, k_{iz,0}, k_{fz,0}\right) = F_1 + r_{0,1}^{f} F_2 + r_{0,1}^{i} F_3 + r_{0,1}^{i} r_{0,1}^{f} F_4 \tag{28}$$

with $F_1 = F\left(q_y, k_{fz,0} - k_{iz,0}\right)$, $F_2 = F\left(q_y, -k_{fz,0} - k_{iz,0}\right)$, $F_3 = F\left(q_y, k_{fz,0} + k_{iz,0}\right)$ and $F_4 = F\left(q_y, -k_{fz,0} + k_{iz,0}\right)$.

Form factors calculated for various geometries are depicted in Fig. 12. Obviously, characteristic structural properties of the inhomogeneities are reflected in great detail in the GISAXS patterns. In a similar way, it is possible to introduce structure factors, which lead to additional interferences, and hence more complicated scattering patterns.

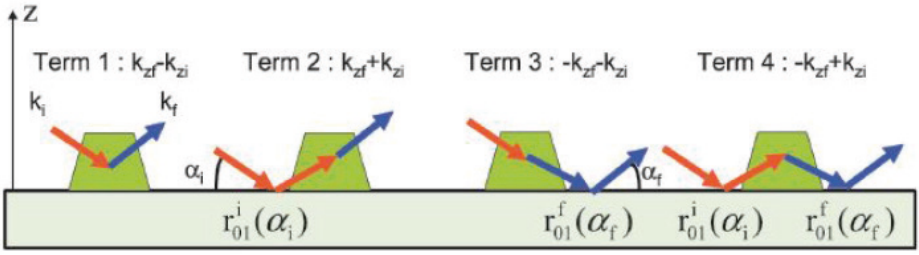


Fig. 11: *The four different contributions to the effective form factor in the DWBA description of GISAXS experiments. The terms r_{01} refer to Fresnel coefficients, Eq. (25).*

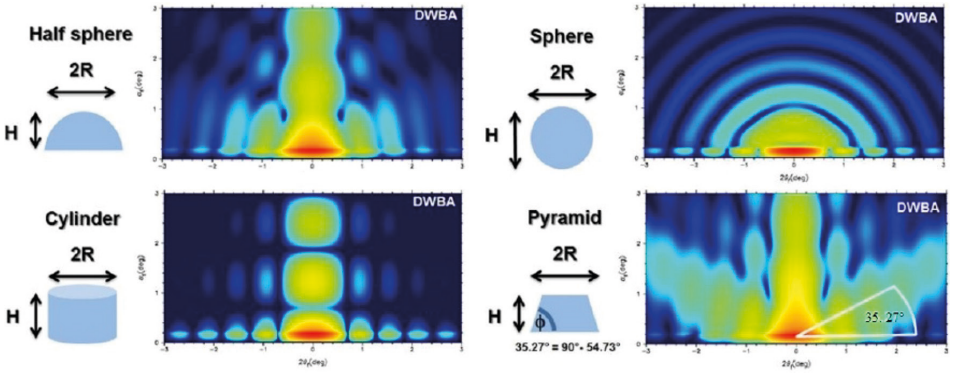


Fig. 12: *Simulation of form factors for various geometries. In view of Eq. (27), $|F(\boldsymbol{Q})|^2$ is plotted for comparison with experimental data.*

GISAXS has been employed to study filamentary switching in memristive materials. In the case of $SrTiO_3$ (STO), the Ti ions that show a valence change are known to agglomerate along line dislocations [14]. This provides an electron density contrast of about 10 - 20 % with respect to the surrounding matrix, sufficient to be detected by small angle X-ray scattering. However, other scattering contributions (interfacial roughness, high-Z electrodes) easily mask the signal, requiring optimized samples for such investigations.
Results from GISAXS experiments at DESY microfocus beamlines are summarized in Fig. 13, together with a sketch of conductive filaments that form as a result of voltage-induced changes in the oxygen stoichiometry. Simulations indicate filament diameters between 10 nm and several 100 nm, depending on the top electrode material [15].

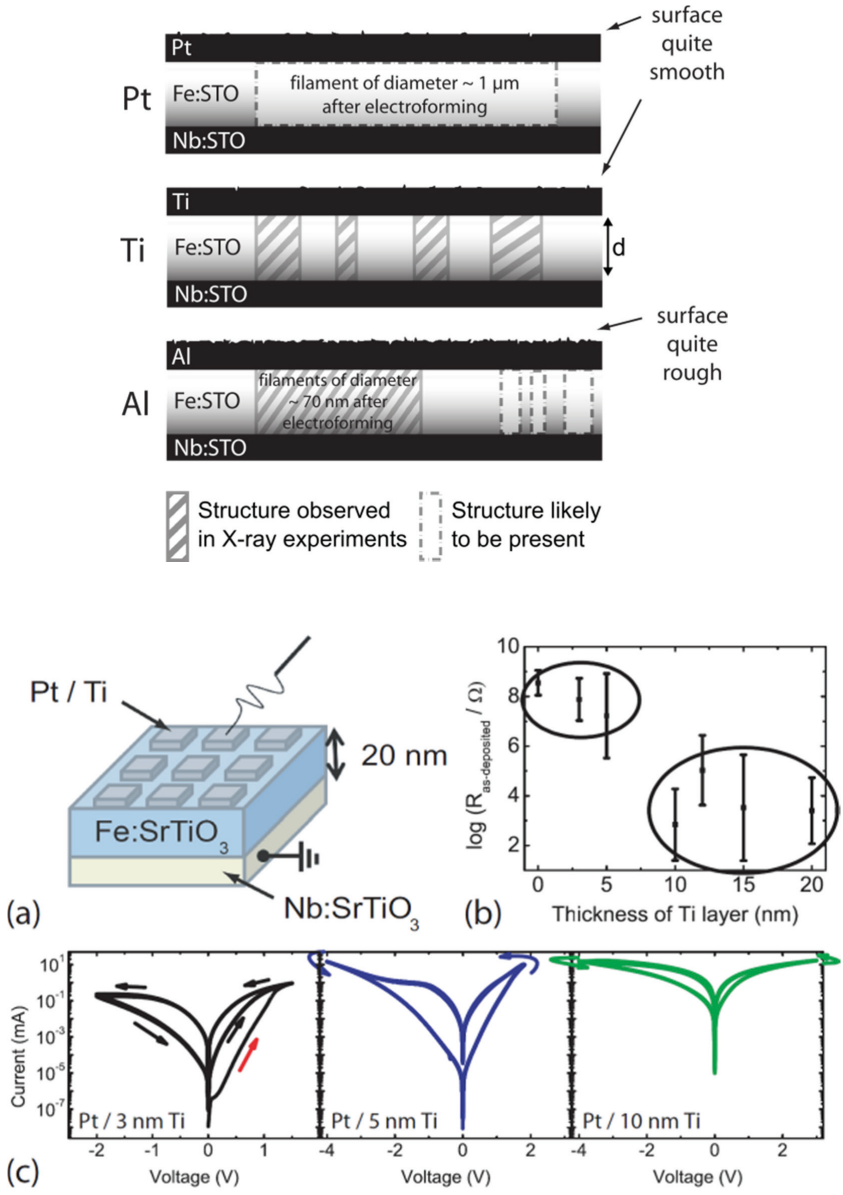


Fig. 13: *Top: Summary of structural results obtained for samples with an active Fe-doped STO layer of 20 nm thickness. Filaments are depicted only schematically. Bottom: Sketch of the sample geometry and I-V curves of metal-insulator-metal structures with different electrode thicknesses [16].*

GISAXS measurements of samples with Ti electrodes show clear signatures of filamentary structures in the insulating STO layer. The resulting GISAXS pattern of an as-deposited sample is presented in Fig. 14 (a). A vertical cut along $q_y = 0.2$ nm^{-1} is presented in Fig. 14 (b). The Yoneda peak at $q_z \approx 0.75$ nm^{-1} is followed by an oscillating intensity, indicated by arrows. From its periodicity of $\Delta q_z = (0.29 \pm 0.03)$ nm^{-1}, the dimension of the scatterers can be calculated to be (21.7 ± 2.2) nm in growth direction. A lateral cut along $q_z = 0.95$ nm^{-1} (Fig. 14 (c)) yields a correlation maximum related to the typical lateral distance between the scatterers.

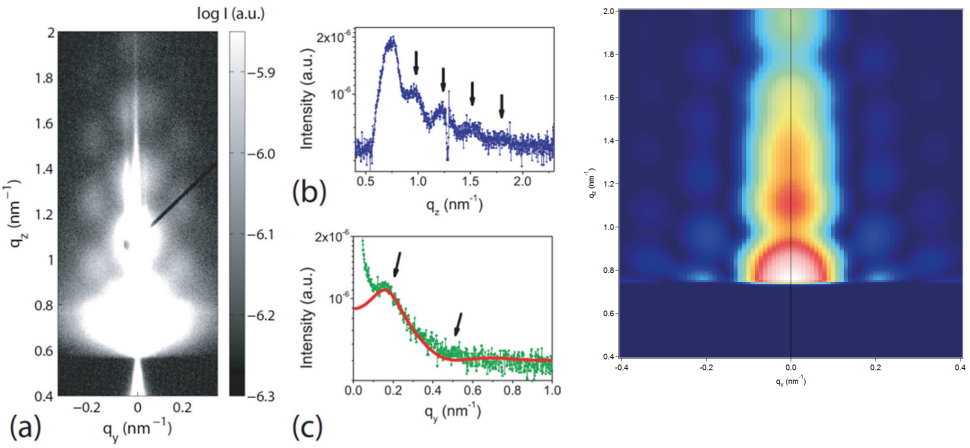


Fig. 14: *(a) GISAXS pattern for a sample with the layer sequence 5 nm Ti / 20 nm Fe-doped STO / Nb-doped STO, with filamentary structures resulting in distinct side lobes. (b) Vertical cut along $q_y = 0.2$ nm^{-1}. (c) Lateral cut along $q_z = 0.95$ nm^{-1}. Red line: simulation based on a cylindrical model [16]. (d) 2D simulation based on a truncated cone structure. The tapered geometry results in vertical shifts of the lobes at $q_y \neq 0$ with respect to those at $q_y = 0$, as is experimentally observed [17].*

Future experiments are envisaged to exploit X-ray beams focused down to the nanoscale in order to obtain lateral resolution despite the large footprint of the beam under grazing angles, as well as in-situ switching of memristive elements while being monitored by GISAXS.

References

[1] R. Waser, R. Dittmann, G. Staikov, K. Szot, Adv. Mat. **21**, 2632 (2009).

[2] J. Als-Nielsen, D. McMorrow, *Elements of Modern X-ray Physics*, Wiley (2011).

[3] H. Frielinghaus, *Structure of Soft Matter: Small Angle Scattering*, Lecture Notes of the 38th IFF Spring School "Probing the Nanoworld – Microscopies, Scattering and Spectroscopies", Forschungszentrum Jülich (2007).

[4] C. Krywka, H. Neubauer, M. Priebe, T. Salditt, J. Keckes, A. Buffet, S.V. Roth, R. Doehrmann, and M. Mueller, J. Appl. Cryst. **45**, 85 (2012).

[5] D.J. Griffiths, *Introduction to Quantum Mechanics*, Prentice Hall (2004).

[6] R. Zeller, *Interaction of Radiation with Matter*, Lecture Notes of the 38th IFF Spring School "Probing the Nanoworld – Microscopies, Scattering and Spectroscopies", Forschungszentrum Jülich (2007).

[7] N.W. Ashcroft, N.D. Mermin, *Solid State Physics*, Oxford Univ. Press (1976).

[8] H.-G. Haubold, *Kleinwinkel- und diffuse Streuung von Röntgenstrahlung*, Lecture Notes of the 18th IFF Spring School "Synchrotronstrahlung in der Festkörperforschung ", Forschungszentrum Jülich (1987).

[9] A. Guinier, G. Fournet, *Small-Angle Scattering of X-rays*, Wiley (1955).

[10] M. Servos, *Bestimmung der Größenverteilung von Partikeln unterschiedlicher Geometrien aus Röntgen-Kleinwinkelstreuung,* Diplomarbeit im Fach Physik, RWTH Aachen University (2011).

[11] L. Spiess, G. Teichert, R. Schwarzer, H. Behnken, and C. Genzel, *Moderne Röntgenbeugung*, Vieweg und Teubner (2009).

[12] S.K. Sinha, E.B. Sirota, S. Garoff, and H.B. Stanley, Phys. Rev. B **38**, 2297 (1988).

[13] G. Renaud, R. Lazzari, and F. Leroy, Surface Science Reports **64**, 255 (2009).

[14] K. Szot, W. Speier, G. Bihlmayer, R. Waser, Nature Materials **5**, 312 (2006).

[15] S. Stille, C. Baeumer, S. Krannich, C. Lenser, R. Dittmann, J. Perlich, S.V. Roth, R. Waser, and U. Klemradt, J. Appl. Phys. **113**, 064509 (2013).

[16] S. Stille, C. Lenser, R. Dittmann, A. Koehl, I. Krug, R. Muenstermann, J. Perlich, C.M. Schneider, U. Klemradt, and R. Waser, Appl. Phys. Lett. **100**, 223503 (2012).

[17] O. Faley, unpublished results.

C3 From Atomic Structure to Properties of Oxides

– Applications of Aberration-corrected TEM

Chun-Lin Jia
Ernst Ruska-Centre for microscopy and spectroscopy with electrons and Peter Grünberg Institute, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

Contents

1 Introduction **2**

2 Quantitative HRTEM based on NCSI technique **2**
2.1 NCSI technique 2
2.2 Iterative procedure for quantitative HRTEM 6

3 Atomic-scale study of electric dipoles across domain walls **10**
3.1 Domain walls in ferroelectric $PbZr_{0.2}Ti_{0.8}O_3$ films 10
3.2 Domain walls in multiferroic $BiFeO_3$ crystal 14

4 The structure and chemistry across a single-unit-cell layer of $LaAlO_3$ embedded in $SrTiO_3$ **17**

5 Summary **19**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

Oxide materials have become increasingly important for electronics applications. In particular, thin films of oxides have been considered as the most promising material basis for various electronic devices such as non-volatile ferroelectric random access memory (FRAM), high-density dynamic random access memory (DRAM), and resistive random access memory (RRAM) [1]. Lattice defects including interfaces, dislocation and local chemical variation have attracted great attentions of research. The electrical properties of these defect areas in most cases show a deviation from the matrix bulk. These unexpected properties can be considered for application in devices for novel functions. Hetero-interfaces and dislocations in oxide systems and domain walls in ferroic materials are particularly interesting since these lattice defects can be engineered by thin film technology and their properties and corresponding structure feature can be tested and investigated by various techniques.

Transmission electron microscopy (TEM) has proven to be a powerful tool for structural characterization of materials. In particular, in the recent decade great progress in the technique of high-resolution transmission electron microscopy (HRTEM) has been made by the successful introduction of the spherical aberration (C_S) correctors [2]. Based on the C_S-corrected microscope point resolution of sub-Angstrom has been achieved. For crystalline materials aberration-corrected microscopy can be used for determining the position of atomic columns with a precision of a few picometres and for determining the chemical occupancy of atomic columns with the precision of a few atomic percent. With quantitative evaluation of image contrast of thin crystal the number of atoms within the atomic columns parallel to the viewing direction has been determined and it is also possible to determine three dimension shape of nano-scale crystal with single-atom precision [3].

In comparison, other structure characterization techniques, such as x-ray and neutron scattering, which are reciprocal-space techniques, provide averaged real-space information from macroscopic areas of material samples. TEM can reveal the structural information from micro-scale to atomic scale. Therefore, TEM and HRTEM are desired techniques for studying the real structural feature at defect-affected areas of matter with sub-Angstrom resolution.

In the present lecture, we focus our discussion on quantitative HRTEM based on negative C_S imaging (NCSI) technique [4,5] and its applications to studying oxide materials.

2 Quantitative HRTEM based on NCSI

2.1 NCSI technique

Here we introduce an imaging technique based on C_S-corrected TEM, the NCSI technique [4,5], which results in an enhanced contrast of image in comparison with conventional positive C_S imaging (PCSI) technique. In order to understand the contrast enhancement under the NCSI condition, an approximation for the object, weak phase object (WPO), has to be used. We note that in reality the sample thicknesses (approx. 2–10 nm for oxides), the WPO is inadequate and only a fully dynamical treatment of the electron scattering and imaging problem can give us an adequate description. In the very thin specimen the phase of the propagating electron wave is modified by the object potential. At the exit plane the wave function can be written in terms of the specimen potential U based on the WPO:

$$\psi_{EP}(r) \approx \psi_0[1 + i\pi\lambda U(r)t] \quad (1)$$

where λ is the wave length and t is the thickness. The wave propagating through the objective lens suffers additional phase shift due to the lens aberrations, in which the spherical aberration and defocus are the chief parameters. Using different C_S-defocus combination we can tune the phase shift of the scattered wave by tuning the aberration function, $\chi(g) = \frac{1}{2}Z\lambda g^2 + \frac{1}{4}C_S\lambda^3 g^4$. The classical imaging condition (PCSI) for optimum phase contrast is obtained by combination of a positive value of C_S with an underfocus, which results in a dark-atom contrast. The imaging condition with a negative value of C_S cooperating with an overfocus (NCSI) leads to a bright-atom contrast.

On the basis of equation (1) the contrast enhancement under the NCSI condition can be discussed as following. At the image plane the wave function can be written as:

$$\psi_{IM}(r) \approx \psi_0[1 \mp \pi\lambda U(r)t] \quad (2)$$

if the objective lens adds a phase of $\frac{\pi}{2}$ or $-\frac{\pi}{2}$ to the diffracted wave, resulting in the image intensity

$$I = [\psi_{IM}(r)]^2 \approx 1 \mp 2\pi\lambda U(r)t + [\pi\lambda U(r)t]^2 \quad (3)$$

which is correct to the second order in $U(r)$. Under the NCSI imaging condition, the sign of the linear term of equation (3) is positive and thus the linear and nonlinear terms are additive. In contrast, for the PCSI condition the sign of the linear term of equation (3) is negative and thus the linear and nonlinear terms are subtractive.

Clearly, the contrast modulation due to the projected potential $U(r)$ is higher when the linear and nonlinear terms are additive than when those are subtractive. We should note that this treatment is only valid for very thin specimens. In experiment, the thickness of used oxide samples is 2-10 nm, for which the weak-phase approximation is no more valid. Therefore, a full dynamical calculation of electron scattering and the contrast transfer under partially coherent illumination is required in order to fully investigate the enhancement of the image contrast under the NCSI condition [6,7].

Figure 1 shows two simulated image of [110] $SrTiO_3$ (STO) under the NCSI condition (a) and PCSI condition (b) for a sample thickness of 3.3 nm. The displayed atom symbols clarify that the NCSI mode leads to a bright atom contrast under a darker background. In contrast, the traditional PCSI mode results in a dark atom contrast. The difference in contrast between the NCSI and the PCSI conditions is already evident from the visual inspection of the two images. For a quantitative comparison, the images are normalized to a mean intensity of one, such that the standard deviation of the intensity reflects the image contrast. For the thickness range of 3.3–6.6 nm, which is most suitable for HRTEM investigations, the image contrast resulting from the NCSI mode is on average by about a factor of 2 larger than the related PCSI contrast [6].

For quantitative HRTEM a key requirement is the good signal-to-noise ratio of the images, which determines the precision of position and occupancy of atomic columns in real structure. Figure 2 shows plots of the image intensity profiles for the three types of columns, SrO (a), Ti, and O (b) from the images shown in figure 1. The blue line profiles were obtained from the image calculated under the NCSI condition and the red line profiles from the image calculated under the PCSI condition. The mean intensity $I_{mean} = 1$ is denoted by a black line. In comparison with the red line profiles the blue line profiles show much higher signal intensity at atomic column positions.

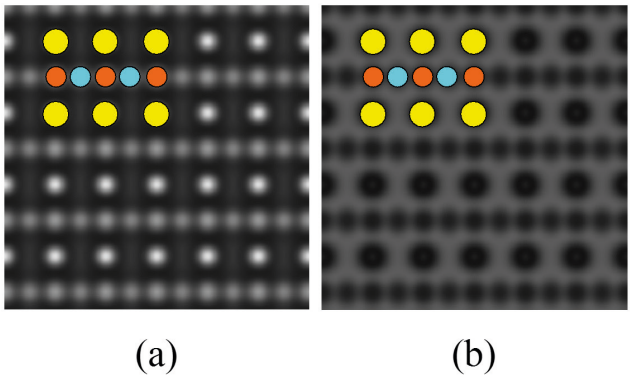


Fig. 1. *Simulated images of STO viewed along the [110] direction (a) under the NCSI condition with C_S = –15 μm, defocus Z = +6 nm, and (b) the PCSI condition with C_S = +15 μm, defocus Z = – 6 nm for a sample thickness of 3.3 nm.*

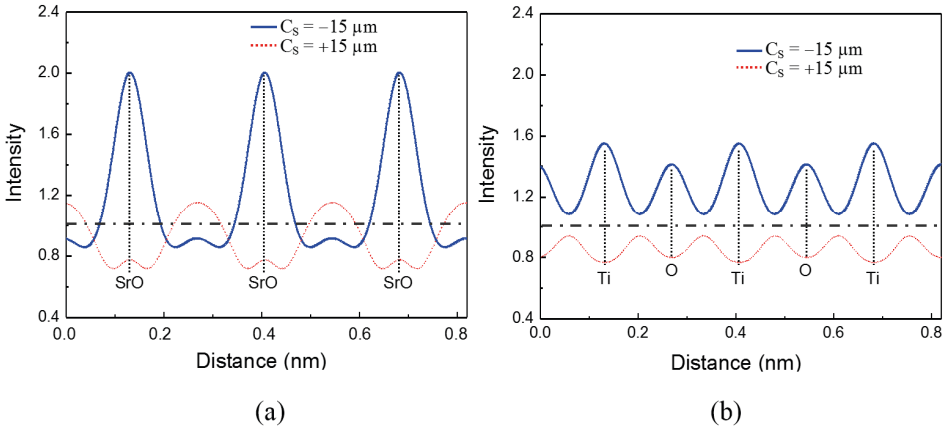


Fig. 2. *Profiles of image intensity for atomic columns of (a) SrO, and (b) Ti and O columns from the NCSI image shown in figure 1a (blue lines) and from PCSI image shown in figure 1b (red lines). Image intensity is normalized to unit mean value.*

Based on the images of the STO crystal shown in figure 1 the effect of an amorphous layer on the measurement precision of column positions was investigated. One up to five random phase object images, which represent the amorphous layers, were added to the images and for each thickness of the amorphous layer the precision of the position measurement was quantified by a peak optimization procedure. Figure 3 shows the precision for determination of the column positions as a function of the number of amorphous layers. Overall, the measurement

precision obtained under the NCSI condition is by a factor of 2 to 3 better than that under the PCSI condition. Under the NCSI condition a precision well below 10 pm is easily obtained even for light-element O columns.

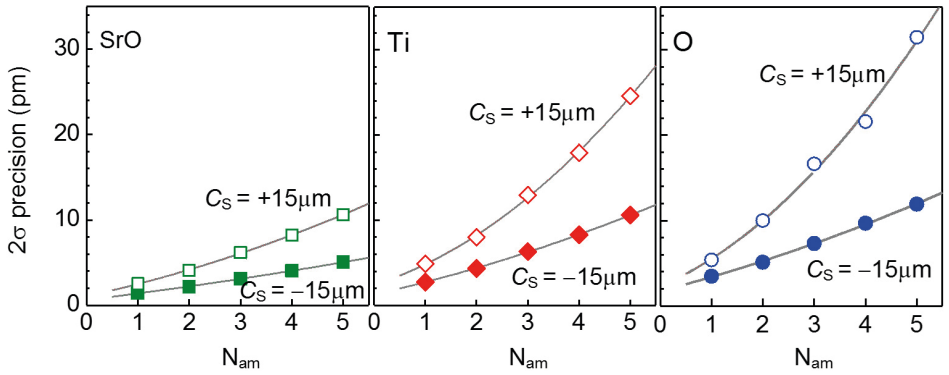


Fig. 3. *Measurement precision for the position of SrO, Ti, and O atomic columns as a function of thickness of amorphous cover layers expressed by the number of the cover layers. The precision data were obtained from the images shown in figure 1 under the PCSI condition (open symbols) and the NCSI condition (solid symbols).*

The dependence of the signal intensity value on the atomic number accumulated in individual atomic column was investigated. Figure 4 shows the results for the atomic columns along the [110] direction of STO. In an unit cell period along the [110] direction of STO the oxygen column includes two oxygen atoms, the Ti column one Ti atom, and the SrO column one Sr atom plus one oxygen atom. Therefore, the sum of atomic numbers over a single unit cell period is 16 (2x8), 22, and 46 (8+38) for the fully occupied oxygen column, the Ti column, and the SrO column, respectively. Under the NCSI condition, the image intensity for all columns follows essentially a linear dependence on the sum of the atomic numbers up to a value of 276. In the case of the PCSI mode, the linearity between the intensity and the accumulated atomic number is already lost for all column types at a value of 100. Most importantly, the linear dependence of the column intensity on the accumulated atomic number is more sensitive (steeper slope) under the NCSI condition than under the PCSI condition.

Based on the image simulations, we have demonstrated that the images obtained under the NCSI condition show great advantages with respect to image contrast, signal intensity for atomic columns, and the linear dependence of the intensity on the atomic number in atom columns. The special features of the negative C_S images are the result of an enhancing combination of phase contrast and amplitude contrast [6,7]. Therefore, the NCSI technique provides optimum condition for direct atomic imaging of material structures, which is the basis for quantitative determination of atomic structures in aberration-corrected transmission electron microscopy.

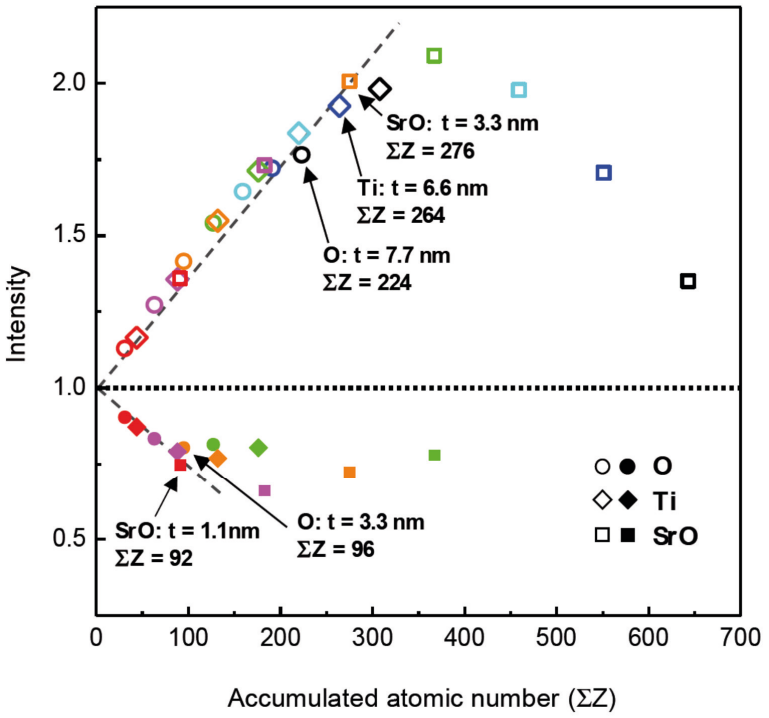


Fig. 4. *Image intensity as a function of the sum of atomic number in atomic columns along the [110] direction of STO for different sample thicknesses under the NCSI condition (open symbols) and under the PCSI condition (solid symbols). The different colours are for different thickness of atomic columns.*

2.2 Iterative procedure for quantitative HRTEM

Under the NCSI condition, the positions of the observed intensity maxima represent already quite well the actual positions of atomic columns. Likewise, for a thin specimen the height of the intensity maxima is roughly proportional to the accumulated atomic charge number along an atomic column. However, it should be noted that in real cases residual lens aberrations and small (unavoidable) tilts of the specimen orientation away from the fully symmetric Laue orientation affect also the contrast of atomic-resolution images. This means that the images recorded in the microscope include not only the structure information but also artefacts induced by the imaging process. In addition, a linear relationship between the observed peak intensity and the actual atomic column occupation is not guaranteed due to the nonlinear nature of electron diffraction. Therefore, in most cases the data measured directly in an HRTEM image cannot simply be used as the real atomic feature for interpretation of various properties of materials. Quantitative comparison between the experimental and the simulated images is the most accurate route for removing these artefacts and thus precisely determining the structure of materials at atomic scale.

In practice, an iterative procedure for image comparison is used, as schematically shown in figure 5, for determining the true atomic structure of material. In the procedure, a primary atomic model for the image area is proposed on the basis of the positions of intensity peaks determined by fitting a two-dimensional Gaussian function to the intensity distribution around the peaks. Using this structure model images are simulated taking the imaging parameters and some sample parameters (e.g. thickness and crystal tilt) as input variables. The imaging parameters can be optimized and estimated by means of evaluating the azimuth tableau of an amorphous area close to the interesting area of sample before the atomic resolution images are recorded. For the NCSI condition with a resolution of 0.08 nm (FEI Titan microscope), the image are recorded using a defocus value $\Delta f = +2 \sim +6$ nm and spherical aberration $C_S = -13$ µm. The residual lens aberrations are usually adjusted to be below certain values, e.g. two-fold astigmatism $A2 < 2$ nm, three-fold astigmatism $A3 < 20$ nm, and axis coma $B2 < 20$ nm.

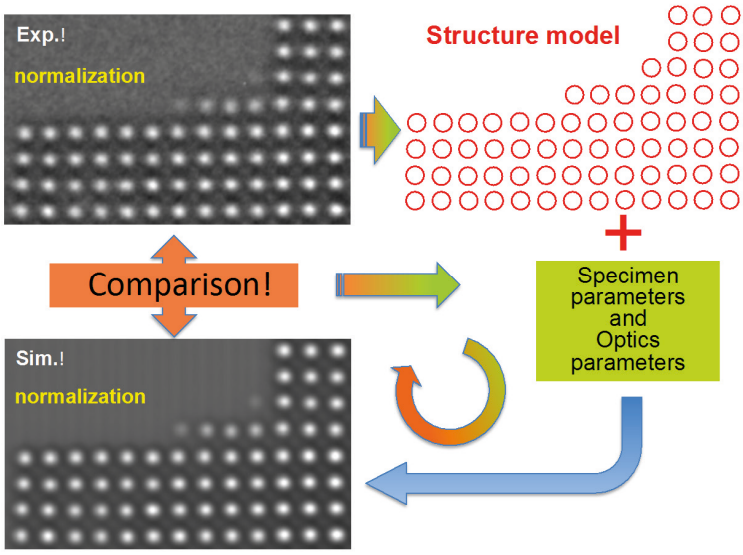


Fig. 5. *Schema of an iterative procedure for quantitative comparison between experimental and simulated images for determining the true atomic structure of material.*

An additional problem is the frequently observed systematic mismatch of the magnitude of the image contrast between simulation and experiment. For solving this problem, in image simulation we need to taken into account the dampening effects on image contrast, which are induced by the modulation transfer function (MTF) of the used charge-coupled-device (CCD) camera [8] and additional image contrast spread function [3,9]. Based on all of the parameters that can be considered, the HRTEM images are simulated and compared quantitatively with the experimental image in an iterative way so that the best match between the simulated and experimental images is obtained by adjusting the input parameters. Only in the case of the best fit between the experimental and the simulated images with respect to the positions and the intensity values of the peaks as well as the true value of the image contrast, one can conclude that the structure model underlying the simulation represents indeed the actual atomic structure.

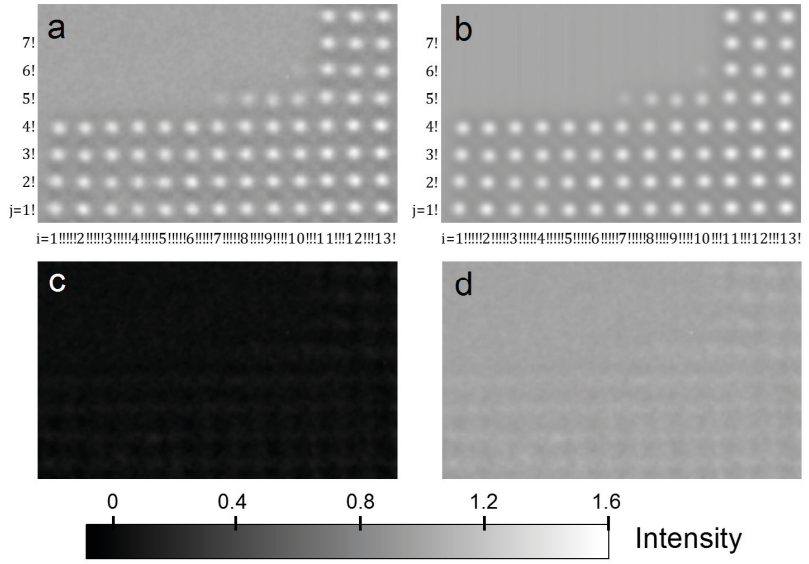


Fig. 6. *Comparison of true contrast between the experimental (a) and the best fitting simulated (a) images of MgO. i and j index the intensity maxima, which were quantified with respect to absolute intensity and geometric position. (c) The difference image between the experimental (a) and the best fitting simulated (b) images. (d) The difference image with enhanced intensity by adding the value of the mean intensity of the normalized simulated image. Note that all of the images are displayed with the same intensity scale.*

An excellent example for quantitative HRTEM is the work on determination of three-dimension shape of MgO nano-scale crystal with atomic resolution [3]. In that work, a complete quantification of experimental and simulated image was performed with an accurate calibration of the relationship between a given atom column and the resulting image intensity. Figure 6a shows an atomic-resolution image of a MgO single crystal specimen containing side terraces parallel to viewing direction (edge of image). The image was recorded along the [100] direction using the NCSI technique. Under the NCSI condition and at the particular specimen thickness the atomic columns, which include the Mg and O atoms stacking alternatively along the [100] direction, appear bright under a dark background. A remarkable feature observed in the original image is the sharp image intensity peaks, amorphous-free, and low noise. Figure 6b displays the simulated image, which exhibits the best match to the experimentally observed image of figure 6a after the application of the iterative comparison procedure. Figure 6c shows the difference image between the experimental and the best match simulated image. In order to show the contrast details, figure 6d displays the difference image, which intensity is artificially enhanced by adding the value of intensity mean of the normalized simulated image. All of the three images are displayed with the same intensity scale. The difference image shows a very low intensity (figure 6c) and very low contrast (figure 6d), indicating the excellent fit between the simulated and the experimental images.

Figure 7a shows a quantitative comparison of the peak intensity at atomic columns measured directly from the experimental image (solid circles) with those of the best fitting simulated image (open squares). The visible discrepancy in some of the circle-square pairs corresponds to an intensity level of image noise in vacuum. Figure 7b plots the difference (δ□□□) in positions (x, y) of the intensity maxima between experimental and simulated images. The standard deviation for the position fitting is 0.5 pm for both x and y coordinates. Based on the data of quantitative comparison the excellent reproduction of the experimental data by the best fitting simulation indicates that the simulation parameters including specimen thickness, specimen tilt, and optical aberrations, have been determined with a sufficiently high accuracy in order to establish a reliable basis for the precise quantification of the atomic structure of the crystal.

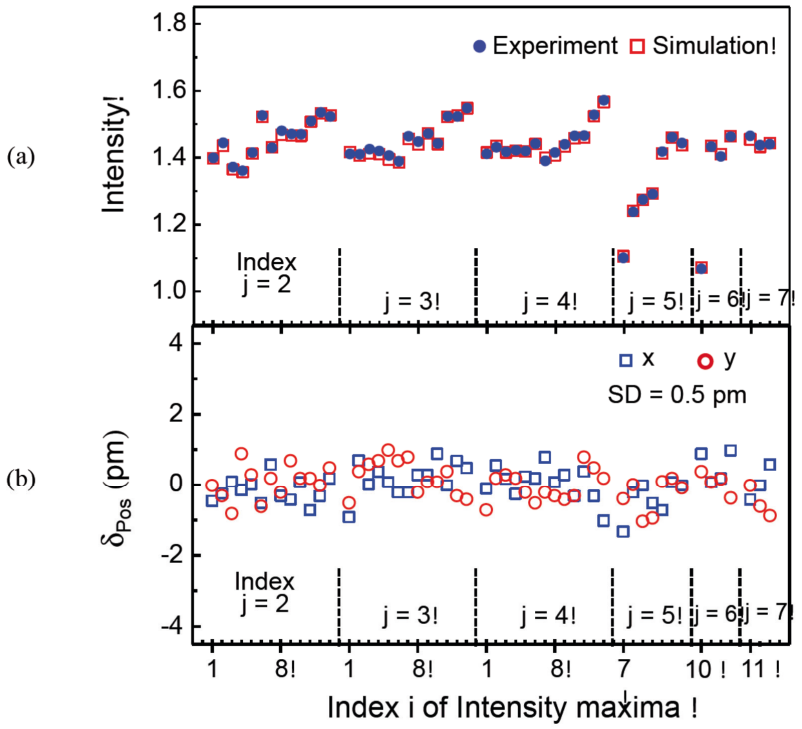


Fig. 7. *(a) Comparison of the data derived directly from the experimental image shown in figure 6a (solid circles) with those derived from the best fitting simulated image shown in figure 6b (open squares) for the peak intensity at atomic positions. (b) The difference δ□□□ in positions of the intensity maxima between the experimental and the simulated images for x (squares) and y (circles) directions. The standard deviation (SD) for the position difference is 0.5 pm for both x and y directions. The indexes i, j of the intensity maxima are referred to those in figure 6.*

In the following, examples are presented for the application of the quantitative HRTEM based on the NCSI technique to characterization of atomic structure and local properties in oxides.

3 Atomic-scale study of electric dipoles across domain walls

The physical properties and structures of domain walls in ferroelectrics and multiferroics have been studied theoretically and experimentally. It was found that the domain walls possess different properties depending on the details of wall structure. The novel properties at domain walls stimulate great interest in experimentally exploring the structure of domain walls at atomic scale.

3.1 Domain walls in ferroelectric $PbZr_{0.2}Ti_{0.8}O_3$ films

Figure 8a shows the cubic structure of paraelectric $PbZr_{0.2}Ti_{0.8}O_3$ (PZT) at high temperature. The atom arrangement in the cubic cell shows a centre symmetry. The charge centres of anions and cations coincide and just compensate. Upon cooling the structure becomes tetragonal (figure 8b) and the material becomes ferroelectric at about 500° C. Inside the tetragonal unit cell the atoms shift to new positions and the centrosymmetry is lost. In particular the positive and negative charge centres no longer coincide. As a result, an electric dipole is formed, resulting in spontaneous polarization (pointing from net negative to net positive charge). In the high-temperature cubic structure there are six equivalent <100> directions that can be chosen as directions of spontaneous polarization direction of the low temperature ferroelectric phase. This means that six types of spontaneous polarization domains are possible. In general case a multi-domain structure is formed in bulk material. As shown in figure 8c, some of the domains are separated by the walls where the polarization vector turns by about 90°. There is another family of wall where the polarization vector changes by 180°. These 180° domain walls can occur in two forms: Longitudinal domain wall (LDW), where the dipoles have head to head or tail to tail orientation, and transversal domain wall (TDW), where the dipoles show head to tail orientation.

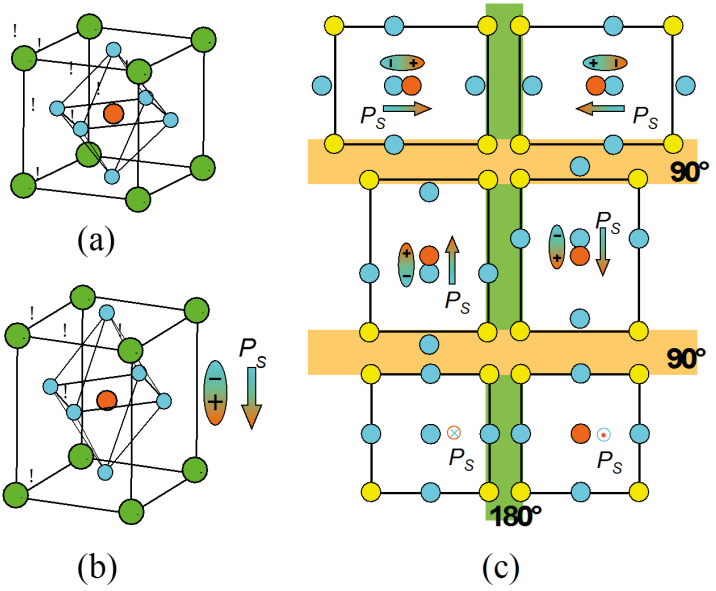


Fig. 8.
(a) The cubic structure of paraelectric PZT at high temperature.
(b) The tetragonal structure of ferroelectric PZT.
(c) Six possible domains and relative domain walls.

Figure 9 shows an HRTEM image of a 10 nm thick PZT layer between two $SrTiO_3$ layers prepared by pulsed laser deposition [10,11]. The image was recorded along the crystallographic [$\bar{1}$10] direction under the NCSI condition. The insets show magnifications of two areas in the upper left side, domain I, and the lower right side, domain II, of the figure. In the insets yellow circles denote PbO atom columns, red circles the Zr/Ti columns and blue circles the oxygen columns. It can be clearly seen from the insets that in domain I the O columns are shifted upward with respect to the neighbouring Zr/Ti columns, while in domain II the shifts of oxygen columns are in the opposite direction. The relative displacements of atoms lead to a separation of the centre of the anionic negative charge of oxygen from that of the cationic positive charge of the metal cations, resulting in spontaneous polarization $\boldsymbol{P}_S$, as indicated by colour arrows. In fact, the image area of figure 9 contains two 180° polarization domains. The position of the respective 180° domain wall is denoted by a dotted line, which was determined directly by mapping the atomic displacements.

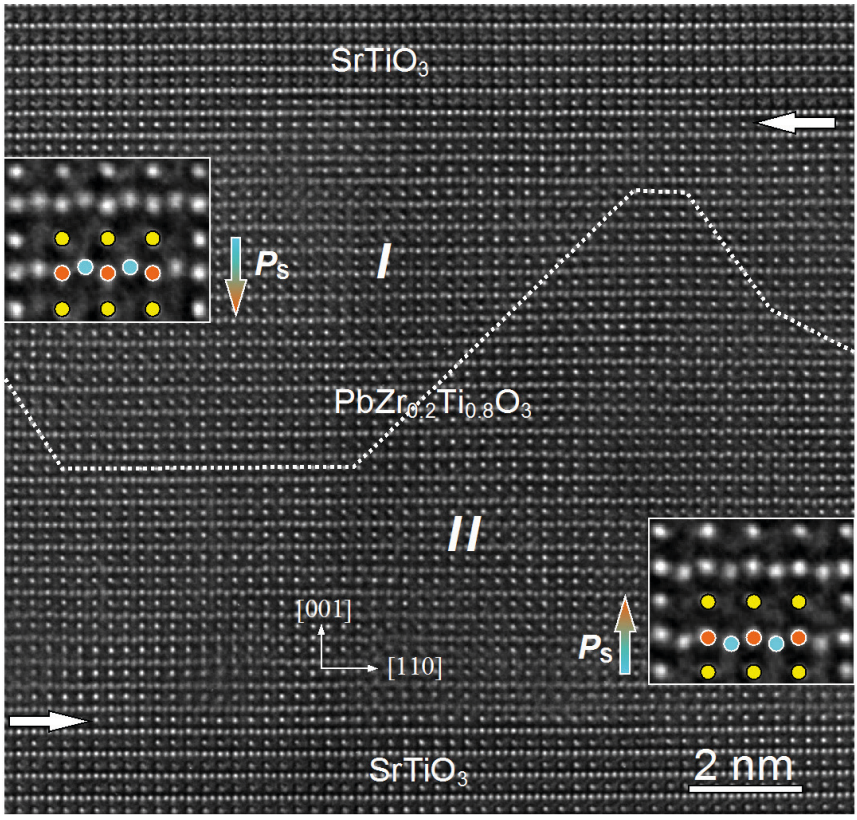


Fig. 9. *Atomic-resolution image of a STO/PZT/STO thin-film heterostructure, recorded along the [-110] direction under the NCSI condition. The horizontal arrows denote the horizontal interfaces between the PZT and the top and the bottom STO film layers. The dotted line traces the 180° domain wall. The arrows "Ps" show the directions of the polarization. Insets display magnifications of the dipoles formed by the displacements of ions in the unit cells (yellow: PbO, red: Zr/Ti, blue: O).*

We are particularly interested in the LDW part of domain wall on the left side of figure 9. This image part of domain wall is enlarged and displayed in figure 10a. The arrows indicate the geometrical centre plane of the wall. We quantified the image area and determined the atom positions using the above-described iterative procedure. Based on the quantitatively determined positions the *c*- and *a*-axis lattice parameter as well as the displacements of the Zr/Ti atomic columns $\delta_{Zr/Ti}$ and O atom columns δ_O were calculated with respect to the PbO column. Since we are only interested in the behaviour of these parameters as a function of distance from the domain wall centre we calculate a mean value for a given distance from the central plane by averaging the position data parallel to the domain wall over the horizontal width of figure 10a.

In figure 10b blue squares and red circles display the off-centre displacements along the [001] direction of the O columns and the Zr/Ti columns, respectively, as a function of the vertical distance from the domain wall plane. Figure 10c shows the spontaneous polarization vs. distance from the central plane of the domain wall. The values of $\boldsymbol{P}_S$ are calculated on the basis of the *c*-axis lattice parameters and the atomic displacements shown in figure 10a and the effective charge values of the ions for $PbTiO_3$. The maximum value of the modulus of $\boldsymbol{P}_S$ is about 75 $\mu C/cm^2$ for domain I and about 80 $\mu C/cm^2$ for domain II. Inside the domain wall area the polarization reaches zero at the wall centre plane and changes direction across the plane.

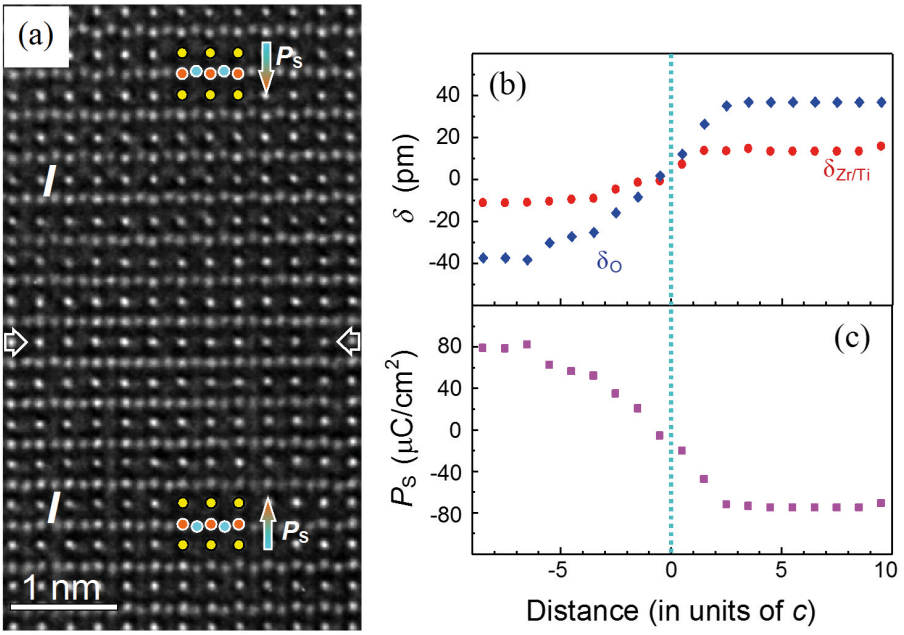


Fig. 10. *(a) Image of an LDW segment. Arrows denote the geometric central plane of the wall, which is referred to as the origin for quantitative analysis of the dipole distortion across the wall area. (b) The displacements of the Zr/Ti atoms ($\delta_{Zr/Ti}$) and the O atoms (δ_O) across the LDW. Positive values denote upward shifts and negative values downward shifts. (c) The spontaneous polarization P_S. The positive values represent upward polarization and the negative values downward polarization.*

Figure 11a shows an atomic resolution image of an area including the interface between the PZT layer and the STO substrate [12], recorded along the [110] direction. In the image the projected unit cell of PZT is schematically indicated in three regions with red circle for the Zr/Ti column, yellow for the PbO column, and blue for the O column. For the sample thickness of about 11 nm, dynamic electron scattering yields a sharp bright contrast for the Zr/Ti and the O atom columns, while the PbO atomic columns are relatively weak. The film-substrate interface was marked by depositing a nominally 1.5 unit cells thick layer of $SrRuO_3$ (SRO) on STO prior to the deposition of PZT. The interface, denoted by a horizontal dashed line, is then determined by observing the plane of RuO_2 serving as a marker.

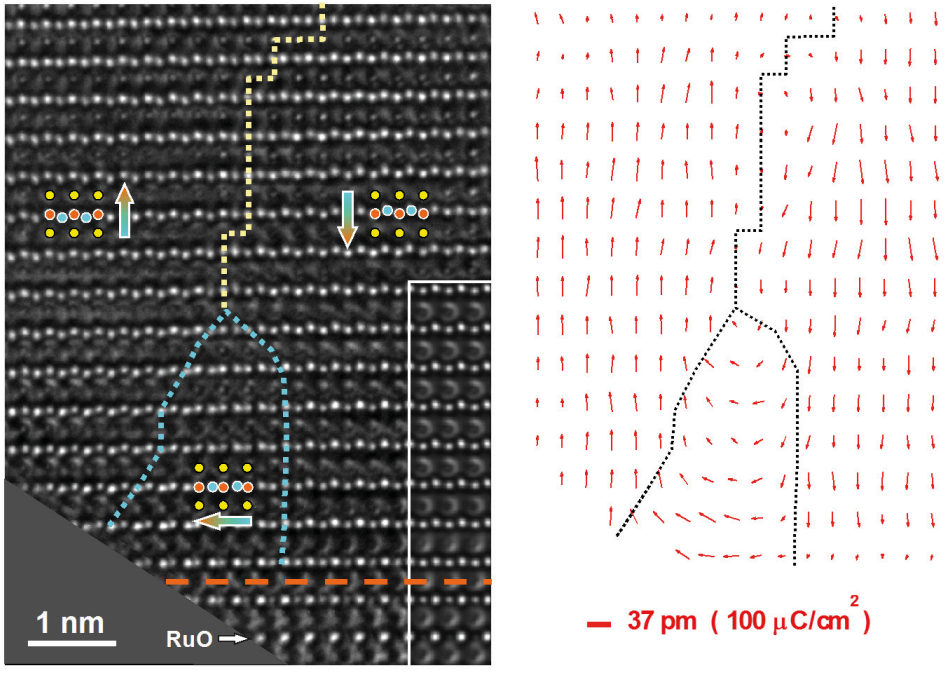


Fig. 11. *(a) Atomic-resolution image of a 180° domain structure in a PZT film close to the interface to the STO substrate, recorded along the [110] direction. The interface is marked by a horizontal dashed line. The domain wall is indicated by a yellow dotted line and the polarization is denoted by arrows. In the centre of the lower half of the image a dotted blue line surrounds an area, where in the centre, the polarization direction makes an angle of 90° with the two large domains. The inset on the right-hand side shows a calculated image demonstrating the excellent match to the experimental image. (b) Map of the displacement vectors for the Zr/Ti atoms (arrows) from the centre of the projected oxygen octahedra. The arrows represent electric dipole moment of unit cell, and thus reveal the continuous rotation of the electric dipoles from "down" (right) to "up" (left) and the structure of electric flux closure.*

In the image of figure 11a the vertical shift of the Zr/Ti positions is clearly visible with respect to the adjacent O positions, indicating a polarized state. In the left-hand part of the image, this shift is upward, while in the right-hand part, it is downward, resulting in the polarization directions indicated by the colour arrows. The opposite direction of the polarization in the two domains forms a 180° domain wall. The position of the domain wall is localized by mapping the atom shifts unit cell by unit cell. At the bottom part of the domain wall, *in-plane* displacements of the Zr/Ti positions with respect to the adjacent oxygen positions are observed between the two 180° domains.

The off-centre displacements of the atoms were determined by the iterative procedure for image comparison based on the image in figure 11a. Figure 11b displays a vector map of atomic displacements. In this map, the middle of arrows is located at the Zr/Ti column positions. The arrows indicate the modulus and the direction of the off-centre displacement with respect to the middle point of the horizontal line connecting the two neighbouring O atom positions. The scale at the bottom left indicates a displacement of 40 pm. We note that a uniform atomic shift of this magnitude corresponds to an integral polarization of 108 $\mu C/cm^2$.

Considering the proportional relation between the off-centre displacement and the electrical dipole moment, the map of displacement vectors provides direct evidence of a continuous rotation of the dipole direction from downward in the right-hand domain through a 90° orientation to upward in the left-hand domain, forming a particular type of flux-closure structure. The reorientation of the dipoles occurs within a well-defined area of triangular shape with the maximum width at the interface of about 2.5 nm. The displacement vector modulus is small at the top, increasing towards the interface. The transition region from the downward orientation of the electric dipoles to the 90° orientation is about two projected unit cell widths on the right-hand side and up to about twice as wide on the left-hand side.

3.2 Domain walls in multiferroic $BiFeO_3$ crystal

$BiFeO_3$ (BFO) is a room-temperature multiferroic material that simultaneously displays ferroelectric and antiferromagnetic properties. BFO has a rhombohedral structure with *R*3*c* space group. It can be derived from the perovskite structure by applying a tensile distortion along the direction of a body diagonal <111> in the pseudocubic notation used here. Along this axis, corner-sharing oxygen octahedra rotate around it in an alternating sense. The cations are displaced from their centrosymmetric positions along [111] inducing spontaneous ferroelectric polarization. In addition, BFO exhibits G-type antiferromagnetic ordering, which is considered to relate to the rotation of the oxygen octahedra. Figure 12 shows a perovskite unit cell of pseudocubic structure in (a), and the projected structure along the [110] direction in (b). From

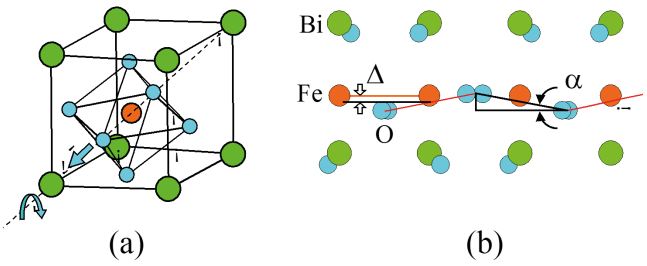


Fig. 12. *(a) Perovskite unit cell of pseudocubic structure of BFO. (b) Projected structure along the [110] direction of the pseudocubic structure.*

the [110] projected structure, the projected off-centre displacement Δ and the projected rotation angle α of oxygen octahedra can be measured [13].

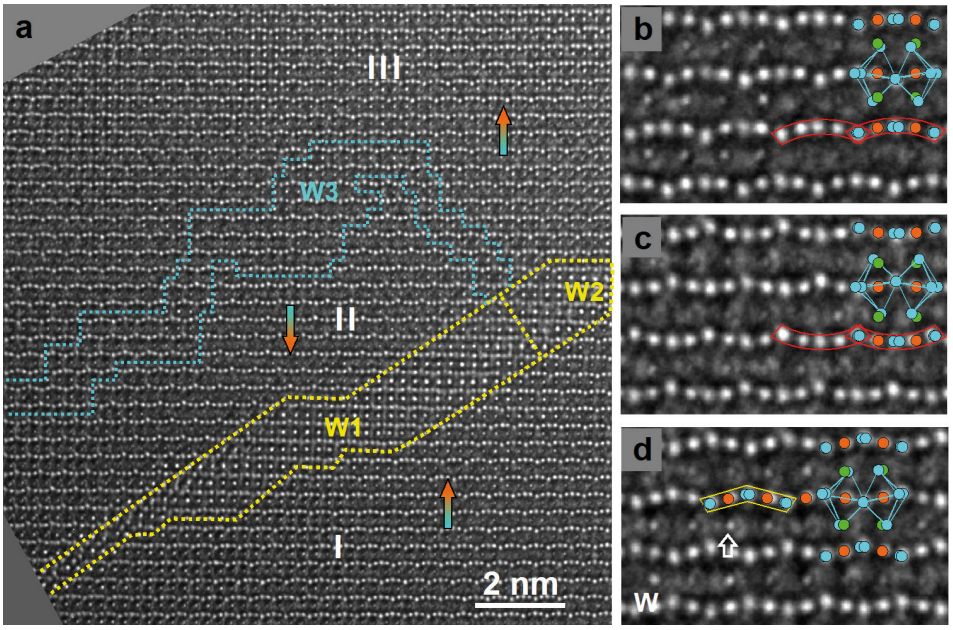


Fig. 13. *(a) Atomic resolution images of domains and domain walls in BFO crystal recorded parallel to the <110> direction. A stripe-domain wall outlined by yellow lines consists of two segments, W1 between domain I and domain II, and W2 between do-main I and domain III. Another domain wall W3 (cyan dotted lines) separates domains II from III. Vertical arrows denote the direction of the <001> component of the <111> type polarization vector. (b) Magnified image of the atom arrangement in domain I. (c) Magnified image of domain II. (d) Magnified image of domain wall area W3.*

Figure 13a shows an atomic-resolution image of a domain structure viewed along the <110> direction, including three domains labelled I, II, and III. The three domains are separated by walls W1, W2, and W3, respectively. In this projection, only the [001] component of the [111] polarization vector can be measured. The domains can be distinguished by checking the off-centre displacement of atoms in each unit cell. Domains I and III exhibit the same projected structure. In the magnified images (figure 13b-d), the structure (one projected unit cell) is indicated. Under NCSI conditions and for the approximately 5 nm thick sample, strong contrast is observed for the Fe and the O atom positions, while that of BiO is weak. The O positions in the images are shifted upward and downward, corresponding to the alternating octahedral rotation. The off-centre displacement of Fe with respect to the middle point of the line connecting two neighbouring O atom positions is clearly visible. This displacement is upward in domains I and III (figure 13b) and downward in domain II (figure 13c). As a result of octahedral rotation and Fe displacement, the chain -O-Fe-O-Fe-O- forms an "arc" inside the projected unit cell. The arc curvature is negative in domains I and III and positive in domain II. In the wall area W3 (figure 13d), the -O-Fe-O-Fe-O- atom positions follow a zigzag line, preserving the rotation of the octahedra while Fe displacements are not seen.

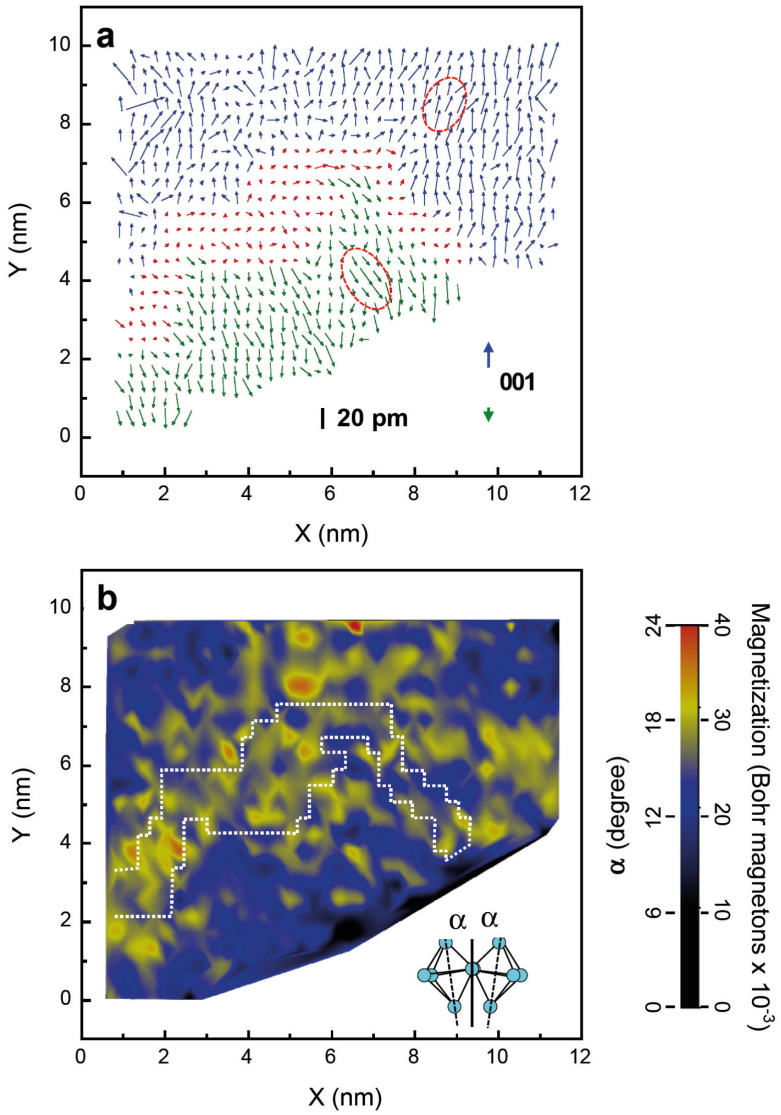


Fig. 14. *(a) Map of the off-centre displacements of Fe atom positions in the domains II (green arrows) and III (blue arrows) and in the domain wall area W3 (red arrows) projected onto the {110} plane. (b) Map of the magnitude of the oxygen octahedron rotation angle, α, projected onto the {110} plane. The area of W3 is indicated by a white dotted line. A scale from blue to red was used for better visualization of the original histogram, which was subsequently smoothed for clarity. Considering the linear relation between the tilting angles of oxygen octahedra and the magnetization, this figure also displays the magnitude of magnetization.*

The off-centre displacements of the Fe atoms and the rotation angles of the O octahedra were studied quantitatively for W3 and the adjoining domains II and III. In measurement of the displacements of the Fe atoms and rotation angles of the oxygen octahedra, the effects of residual lens aberrations and unavoidable small tilt of the crystal have been removed in the iterative procedure for quantitative comparison between experimental and simulated images. Figure 14a shows a map of the Fe displacements projected into the {110} plane. Arrows centred at the Fe positions indicate the magnitude and the direction of the displacement. Inside the domains, the mean value of the displacement component is about 17 pm, which is in good agreement with the value of 19 pm derived from the structure of BFO. In the domain-wall area W3 (red arrows), the displacement changes to the low but finite value of 6 pm. A striking feature in figure 14a is the high degree of disorder on the atomic scale. Both the magnitude and the direction of the projected displacement vector exhibit substantial random deviations from the exact [001] direction. In some areas, nanometre-scale regions with essentially identical directions of dipole vectors can be recognized (red ellipses) showing large deviations (up to a few tens of degrees) from [001].

Figure 14b shows the magnitude of the octahedron rotation angle in colour-coded form. We find that the disorder in the form of fluctuations of the rotation angle extends over the whole image, i.e. both the domain area and the domain-wall area are affected. For the rotation angle projected into the {110} plane, we obtain a mean value of 12.8°±0.14° by averaging over the yellow-dominated area, and 10.7°±0.20° over the blue-dominated areas. Considering the linear relation given in [14] between the tilting angles of oxygen octahedra and the magnetization, this map demonstrates changes of unit cell magnetization from area to area.

4 The structure and chemistry across a single-unit-cell layer of $LaAlO_3$ embedded in $SrTiO_3$

Novel functional properties of the interface between insulating $LaAlO_3$ (LAO) and STO have been measured, including metallic conductivity, superconductivity, and magnetism. These physical phenomena, which are not intrinsic to the bulk material, can be related to the local atomic rearrangements at the interface, the special feature of oxide structure, and the strong correlation of electrons to the ionic lattices. For a comprehensive understanding of the origin of these novel properties, subtle details of the atomic structure at the interface area must be taken into account. A simultaneously quantitative determination of the structural and the chemical details at an identical specimen area has become great challenge of transmission electron microscopy. By means of quantitative HRTEM, which is described in the section 2, the chemistry and structure were investigated on atomic scale near a nominally single-unit-cell layer of LAO, which is sandwiched between a capping STO layer and the STO substrate [9].

Figure 15 shows an atomic-resolution cross-sectional image of a sample containing a nominal single-unit-cell layer of LAO sandwiched between a capping layer of STO and the STO substrate. Under the NCSI condition and at the particular specimen thickness of about 4 nm one obtains a bright contrast for all atom columns, including the oxygen columns. While the contrast of the Ti and the SrO atom columns is relatively strong, the contrast for the LaO atomic columns (marked by an arrow) is comparatively weaker, at a similar level as that of the oxygen columns. The atomic displacements and the chemical intermixing across the nominally single LAO unit cell are iteratively determined by a comparison of the position and the height of the intensity maxima between experimental (figure 15a) with corresponding image simulations (figure 15b).

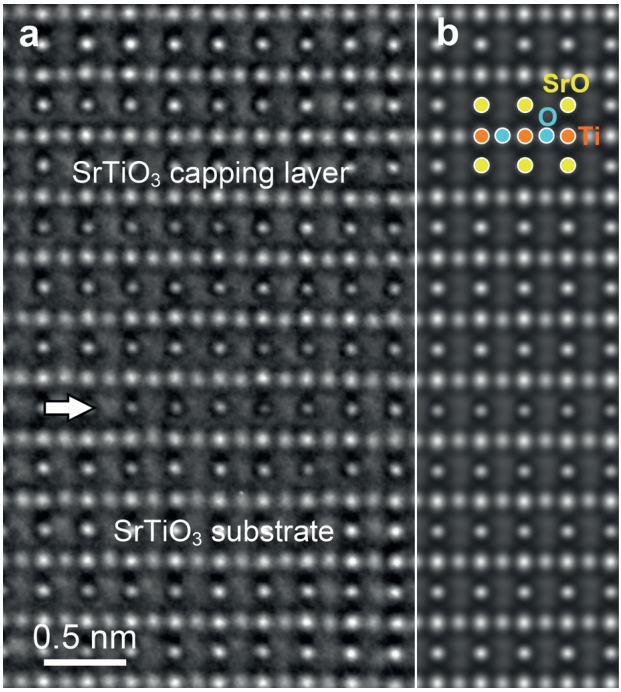


Fig. 15. *(a) Atomic-resolution image of the nominally single unit cell layer of LAO embedded in STO, which was recorded along the [110] direction of STO under the NCSI condition. The arrow denotes the nominally single LaO plane. (b) Simulated image with the best match to the experimental image.*

The final data are displayed in figure 16. As shown in figure 16a, the AO-type columns, SrO and LaO, do not show evident shifts across the single unit cell layer. In contrast, shifts of the oxygen columns are measured in the BO_2 plane (B represents Ti in STO and Al in LAO). The positive values denote the upward shifts and negative ones the downward shifts with respect to the B-type columns (referring to the image of figure 15). On the left (corresponding to above the nominal LaO plane in the image) the shifts are downward, and on the right (below the nominal LaO plane in the image) are upward. All shifts point towards the nominal LaO plane. The collective displacement of oxygen atoms leads to a shift of the oxygen octahedron centres away from the B-type columns, implying a separation of the centre of negatively charged oxygen from the positive charge centre of the cations and thus electric polarization. Figure 16b shows concentration profiles across the nominal LaO plane, which is directly obtained from the final structure model determined by the iterative refinement procedure. The intermixing occupancy of A-site by La and Sr atoms is evident and extends to about 4 unit cells. In the nominal LaO plane only 55% A-sites are occupied by La atoms and the other 45% by Sr atoms. The cation intermixing is stronger in the STO capping layer, extending over three unit cells above the nominal LaO plane in comparison to one unit cell below the nominal LaO plane in the substrate. In the same area, intermixing of Ti and Al occurs also in the B-type columns. In the BO_2 atomic plane direct above the nominal LaO plane, which is nominally expected to be AlO_2 plane, Al atoms occupy only 35% lattice sites. In the plane below the nominal LaO plane, which is the TiO_2 terminating plane of the STO substrate, an intermixing of 30% Al with 70% Ti was determined. In the area of cation intermixing the image intensity values for the oxygen columns show detectable deviation from the substrate area, leading to an oxygen deficiency of about 10%.

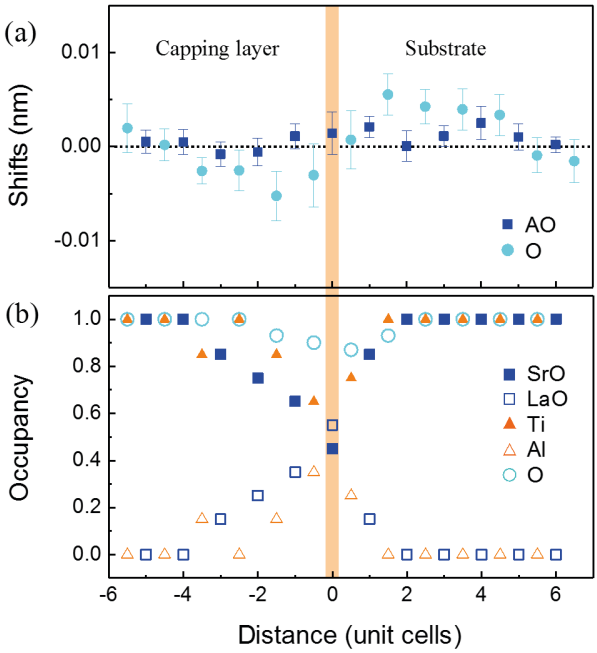


Fig. 16. *(a) The shifts of the AO-type columns (SrO and LaO), and of the oxygen columns across the single unit cell layer. (b) Chemical occupancy. The position of the nominal LaO atomic plane is marked by a vertical thick line.*

5 Summary

In summary, we have demonstrated with several examples that structural details, such as the atom positions and chemical occupancy in atomic columns that parallel to electron beam, can be obtained from a single HRTEM image. By quantitative comparison of image simulations with the experimental images recorded under the NCSI condition, the atom displacements can be measured with a precision of a few picometres. For ferroelectric materials, based on the precisely determined off-centre displacements the electric dipole of unit cell can be calculated and thus the local polarization can be investigated across domain walls and lattice defect areas. For some of magnetic oxides local magnetization can be also studied taking the simple relation to the rotation angle of the oxygen octahedra. By quantitative analysis of the image contrast the intermixing of cations in atomic columns and oxygen deficiency at an interface can be determined. In addition to the above-described examples, another excellent application of the quantitative HRTEM is to determine the three-dimension shape of nano-scale MgO crystal with atomic resolution from a single image [3]. The successful application of the quantitative HRTEM to solving structural problems has played important role in understanding the relations between structure and properties of materials and is expected to make more progress in the future materials research.

Acknowledgments

The author acknowledges the collaborations with K. Urban, A. Thust, J. Barthel, L. Houben. M. Lentzen, S.B. Mi, L. Jin, D. Hesse, M. Alexe, I. Vrejoiu, R. Dittmann and F. Gunkel.

References

[1] R. Waser, R. Dittmann, G. Staikov and K. Szot, Redox-Based Resistive Switching Memories – Nanoionic Mechanisms, Prospects, and Challenges, Adv. Mater. **21**, 2632 (2009).

[2] K. Urban, Studying atomic structures by aberration-corrected transmission electron microscopy. *Science* **321**, 506 (2008).

[3] C. L. Jia, S. B. Mi, J. Barthel, D. W. Wang, R. E. Dunin-Borkowski, K. W. Urban, A. Thust，Determination of the 3D shape of a nanoscale crystal with atomic resolution from a single image, *Nature Mater.* **13**, 1044 (2014).

[4] C.L. Jia, M. Lentzen and K. Urban, Atomic-resolution imaging of oxygen in perovskite ceramics, *Science* **299**, 870 (2003).

[5] C.L. Jia, M. Lentzen and K. Urban, High-resolution transmission electron microscopy using negative spherical aberration, *Microsc. Microanal.* **10**, 174 (2004).

[6] C.L. Jia, L. Houben, A. Thust and J. Barthel, On the benefit of the negative-spherical-aberration imaging technique for quantitative HRTEM, *Ultramicroscopy* **110**, 500 (2010).

[7] K. Urban, C.L. Jia, L. Houben, M. Lentzen, S.B. Mi, K. Tillmann, Negative spherical aberration ultrahigh-resolution imaging in corrected transmission electron microscopy, *Phil Trans R Soc A* **367**, 3735 (2009).

[8] A. Thust, High-resolution transmission electron microscopy on an absolute contrast scale. *Phys. Rev. Lett.* **102**, 220801 (2009).

[9] C.L. Jia *et al.* Atomic-Scale Measurement of Structure and Chemistry of a Single-Unit-Cell Layer of $LaAlO_3$ Embedded in $SrTiO_3$. *Microsc. Microanal.* **19**, 310 (2013).

[10] I. Vrejoiu, *et al.* Intrinsic Ferroelectric properties of strained tetragonal $PbZr_{0.2}Ti_{0.8}O_3$ obtained on layer–by–layer grown, defect–free single–crystalline films. *Adv. Mater.* **18**, 1657 (2006).

[11] C.L. Jia, S.B. Mi, K. Urban, I. Vrejoiu, M. Alexe and D. Hesse, Atomic-scale study of electric dipoles near charged and uncharged domain walls in ferroelectric films, *Nature Mater.* **7**, 57 (2008).

[12] C.L. Jia, K. Urban, M. Alexe, D. Hesse and I. Vrejoiu, Direct observation of continuous electric dipole rotation in flux-closure domains in ferroelectric $Pb(Zr,Ti)O_3$, *Science* **331,** 1421 (2011).

[13] C.L. Jia, L. Jin, D. Wang, S.B. Mi, M. Alexe, D. Hesse, H. Reichlova, X. Marti, L. Bellaiche, K. Urban, "Nanodomains and nanometer-scale disorder in multiferroic bismuth ferrite single crystals", *Acta Materialia* **82,** 356 (2015).

[14] L. Bellaiche, Z. Gui and I.A. Kornev, A simple law governing coupled magnetic orders in perovskites. *J. Phys.: Condens. Matter* **24**, 312201 (2012).

C 4 HRTEM Based Spectroscopy Techniques

C. B. Boothroyd
Ernst Ruska-Centre for Microscopy and spectroscopy with electrons
Forschungszentrum Jülich

Contents

1 Introduction **2**

2 Energy-dispersive X-ray spectroscopy (EDX) **3**

3 Electron energy-loss spectroscopy (EELS) **7**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

High-resolution transmission electron microscopy (HRTEM) and scanning transmission electron microscopy (STEM) are able to provide images showing the positions of atoms in materials. While high-angle annular dark-field (HAADF) STEM images show contrast that is proportional to the atoms atomic number, none of these techniques is able to identify the atoms. HRTEM based spectroscopy techniques make use of the interactions between the incident electron beam and atoms in a material to give a signal that allows identification of the elements present on an atomic scale.

When an incident electron beam strikes a sample, there are a number of signals that are emitted from the sample [1]. These, as shown in figure 1, include light, Auger electrons, secondary electrons, backscattered electrons, X-rays and electrons transmitted through the sample if the sample is thin enough. All these signals are useful in understanding the sample. However the signals most useful for elemental composition determination are X-rays, Auger electrons and the transmitted electrons that have lost energy.

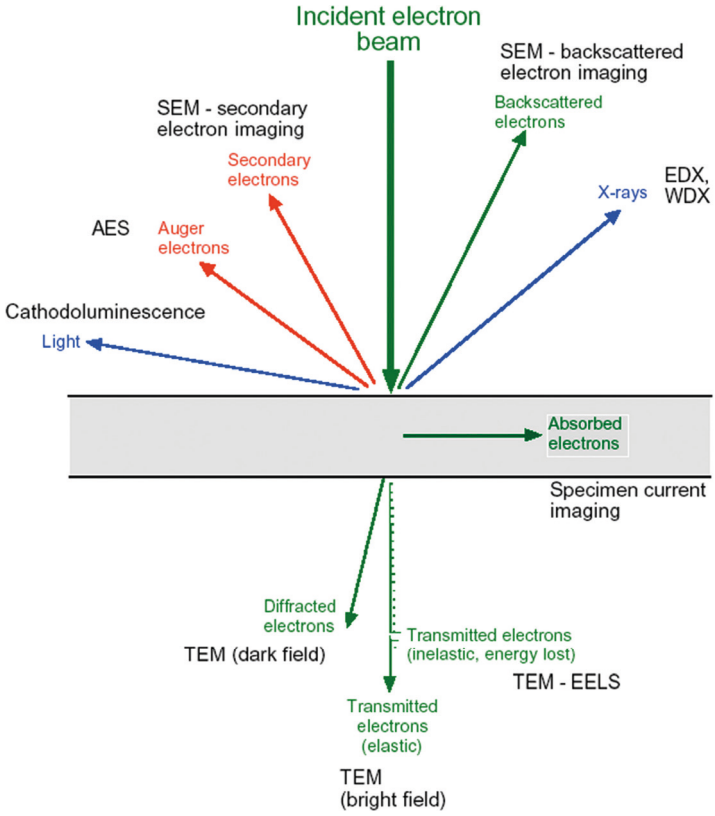


Fig. 1: *Signals emitted when an electron beam is incident on a TEM sample.*

In this chapter we will discuss the most common two of these signals, X-rays as used in energy-dispersive X-ray spectroscopy (EDX or EDS or EDXS) and transmitted energy-loss electrons as used in electron energy-loss spectroscopy (EELS).

X-ray analysis can be used in all types of electron microscopes ie scanning and transmission microscopes (and even focused ion beam microscopes) as X-rays are emitted in all directions from the point where the electron beam hits the sample. Energy loss spectroscopy is only relevant to transmission electron microscopes where the sample is thin enough for a significant proportion of the electron beam to pass through the sample.

2 Energy-dispersive X-ray spectroscopy (EDX)

In EDX we make use of the X-rays emitted when an electron beam passes through a sample as a means of identifying atoms. Although EDX can be used in all types of electron microscope, we will concentrate here on EDX in a high-resolution transmission electron microscope. In this case the sample is thin enough (typically <100nm thick) and the electron beam voltage high enough (typically 200-300kV) that most of the incident electrons pass through the sample.

Interaction volume

When an incident electron beam is focused on the surface of a sample the electrons penetrate the sample and are scattered into an interaction volume [2]. As an example, figure 2 shows the interaction volume for 20kV electrons striking a thick sample of copper. The electrons penetrate to a depth of around 1.5μm and spread to a width of around 1μm, thus limiting the compositional resolution available. It should be noted that the highest electron concentration and thus highest X-ray emission is around the point where the beam hits the sample, so the overall resolution is not as bad as this figure would suggest. For a high-resolution TEM sample, the thickness would typically be around 50nm and the electron energy 200kV or above, so that the loss of resolution due to the interaction volume is much less than for EDX in a SEM and for many samples atomic resolution is possible. Unfortunately as most of the electrons pass through thin samples without interaction, the X-ray signal available from HRTEM samples is much less than for bulk SEM samples with equivalent beam current.

X-rays and X-ray generation

When an electron hits a material X-rays are formed by 2 processes, giving rise to bremsstrahlung and characteristic X-rays.

The emission of bremsstrahlung X-rays is due to the incident electrons being decelerated by the electric field around the atoms in the material. As generated they contain all energies from zero up to the incident beam energy in a continuous spectrum, with the lowest energies having the highest intensity (see figure 3). However the lowest energy bremsstrahlung X-rays are absorbed most strongly on their way out of the sample meaning that the characteristic shape of the observed bremsstrahlung emission has a peak in intensity at around a few kV energy. Bremsstrahlung X-rays are of little use for elemental analysis as their spectrum changes little between different elements.

An incident electron can interact with an atom removing an electron from an inner shell leaving the atom in an excited state (ionised). Later on this ion loses energy as one of the outer shell electrons falls into the inner shell vacancy. This excess energy can be emitted as either an X-ray or an Auger electron. When the energy is emitted as an X-ray the energy is sharply peaked and this gives rise to characteristic X-rays.

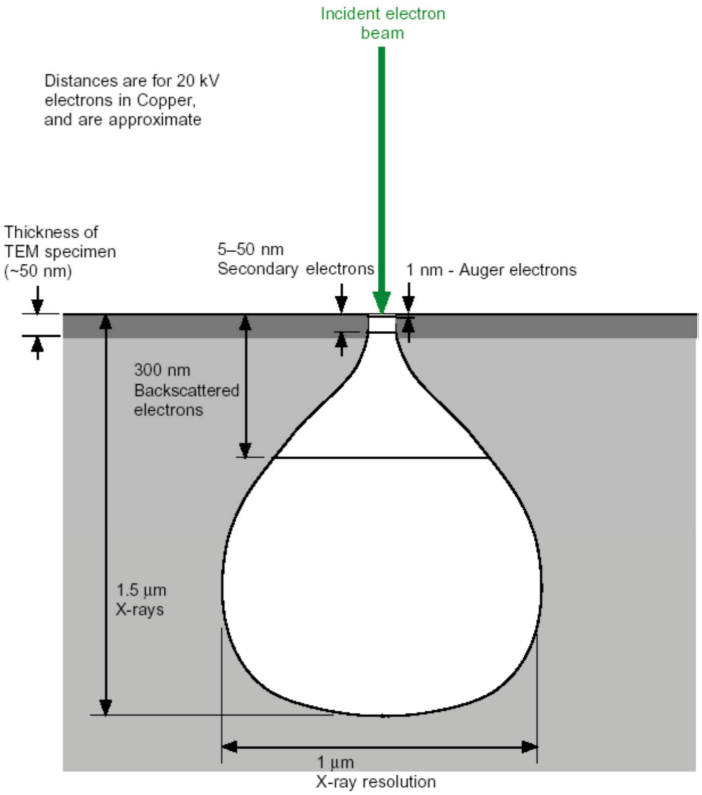


Fig. 2: *Interaction volume for a 20kV electron beam incident on a thick sample of copper. Also shown is the part of the interaction volume in a 50nm thick TEM sample.*

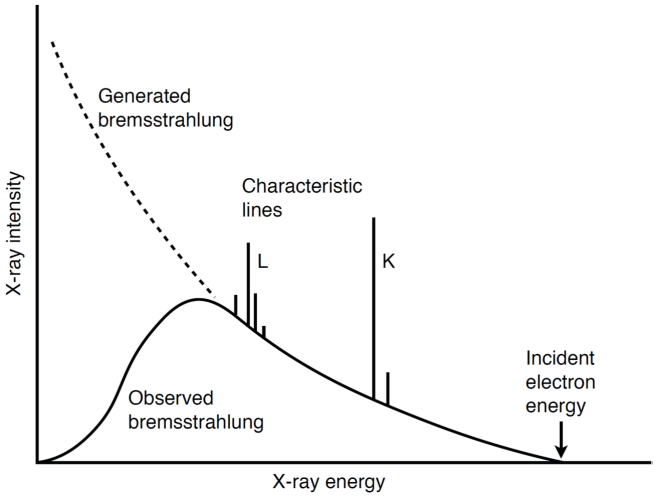


Fig. 3: *Schematic X-ray emission spectrum showing the bremsstrahlung and characteristic X-rays.*

The probability of X-ray emission (rather than Auger emission) is given by the fluorescence yield, ω. The fluorescent yield is small for low atomic number elements meaning that light elements have a low X-ray yield.

Characteristic X-rays are much more useful than bremsstrahlung X-rays as their energy is both sharply peaked and different for each element.

Recording spectra

To record X-ray spectra when an electron beam strikes a sample the energy of each X-ray photon must be measured and their number counted. For modern EDX detectors a silicon detector is used which produces electron-hole pairs whose number is proportional to the energy of each X-ray. Such detectors can measure and count up to around 10^5 X-rays per second but their energy resolution is limited to around 130eV by the leakage current in the detector and by electron-hole pair counting statistics. This means that the major X-ray emission lines for each element can be separated but there can be overlaps at low energies, particularly below 2kV.

Collecting and interpreting EDX spectra

These are the steps required to collect and interpret EDX spectra from an unknown material.

1. Ensure that the incident beam energy is high enough. For quantitative work the beam energy must be greater than twice the highest peak energy of interest otherwise the peak intensity will be adversely affected.
2. For interpretable spectra the count rate must be high enough but not too high to saturate the detector (ie below 50% dead time). It is also often necessary to tilt the specimen to prevent the edge of the specimen holder shadowing the detector.
3. Use prior knowledge of the sample to know which elements are likely to be present and which are unlikely to be present. Work from high energy where there are fewer peaks to low energy identifying peaks and confirm elements by looking for other peaks from the same element.
4. For the energy range 0–20kV (typical of most spectrometers) elements B (Z = 4) to Ru (Z = 44) have a K peak and for Z>16 (S) a K_β will be present with an intensity of about 10% of the K_α peak. There will also be L peaks for Cl (Z = 17) and higher with the Lβ peak and other minor peaks visible for Mo (Z = 42) and above. In addition, Ag (Z = 47) and higher Z elements have an M peak with the highest energy M peak (for U) being at 3.2kV.
5. There are a number of peak overlaps to watch out for, two of the most common are S K (2.31 keV), Mo L (2.29 keV), Pb M (2.35 keV) and N K (0.39 keV), Ti L (0.45 keV).

Artefacts

There are many possible artefacts in EDX spectra, some of which can be mistaken for peaks and it is important to understand their causes and their effects on spectra.

Escape peaks occur when a Si K X-ray escapes from the Si detector. The energy recorded is reduced by the energy of a Si K X-ray. An extra peak appears 1.74 keV below any intense peak. The intensity is about 0.2–2% of the main peak. Escape peaks are most often seen for elements between P K (2.0 keV) and Zn K_α (8.6 keV).

Sum peaks (or coincidence peaks) appear when two X-rays arrive at the same time so that the electronics cannot distinguish them from a single X-ray. Thus an extra peak appears at double the energy of each strong peak. This happens for high count rates and thus high dead times (typically > 50%).

There are many possible sources of stray radiation. X-rays can be collected from various parts of the microscope chamber, detector, sample holder, or parts of the sample away from the area of interest. They can be created due to backscattered electrons and/or X-rays hitting parts of the specimen and/or chamber and exciting secondary X-rays. This is especially a problem for TEM, where there is little space surrounding the specimen. For this reason, it is important to remove the objective aperture, which is particularly good at scattering electrons back onto the specimen. Even then, some secondary fluorescence peaks, such as the Cu peak from a Cu support grid are always present. The presence and strength of secondary fluorescence peaks in TEM EDX spectra mean that TEM EDX can never really be quantitative!

Coherent bremsstrahlung peaks appear as two or three small peaks at low energy, which can be mistaken for low Z elements, e.g. Si or S. They appear in crystalline materials when the beam is at a zone axis.

Quantitative Energy dispersive X-ray spectroscopy

In *qualitative* analysis, we are just interested in finding the elements present in our sample and getting a rough idea of how much of each is present. This can be done just by looking at and identifying the peaks in the X-ray spectrum. In *quantitative* analysis, we want to measure the proportion of each element present in the sample as accurately as possible, by measuring the areas under the X-ray peaks. To perform quantitative analysis, we first need to separate the X-ray counts in each peak from the background and separate overlapping X-ray peaks. Then we need to convert the number of X-ray counts measured from each element to an atomic fraction.

Although it is possible to estimate the area of an X-ray peak by subtracting a background interpolated from adjacent peak free regions, normally some more sophisticated processing is needed to extract reliable peak areas from overlapping peaks. This can be done by background modelling, where the background is calculated using an initial estimate of the sample composition, fitted to regions of the spectrum with no peaks then subtracted to leave just the characteristic peaks. Alternatively Fourier or top hat filtering can be used to remove the slowly varying background. This alters the shape of the spectrum peaks. A multiple least-squares fitting method is then used to decompose the experimental filtered spectrum into fractions of each filtered standard peak.

For EDX in TEMs where specimens are thin and absorption and fluorescence corrections are relatively small, conversion of peak areas to compositions is normally done using the Cliff-Lorimer ratio technique. This relies on the equation $C_A/C_B = k_{AB}(I_A/I_B)$ where A and B are two elements in the unknown alloy. k_{AB} is called the Cliff-Lorimer k factor or just the “k-factor". It is not a constant, but is related to the atomic number correction factor (Z). Note that no standards are needed for this method.

Since measuring k factors for every pair of elements would be difficult, they are measured with respect to one element, usually Si. Si was originally chosen because many minerals contain Si, and its K edge has a high enough energy to be detected on the Be window detectors then available (c.f. oxygen, the other obvious choice). Then $k_{AB} = k_{ASi}/k_{BSi}$.

To work out the absorption and fluorescence corrections (and background modelling), the sample composition needs to be known, hence an iterative approach is needed.

For quantitative analysis to give reliable results the sample must be of uniform composition over the area that the beam spreads to, plus adjacent areas that the X-rays pass through, the sample must be flat and oriented at the expected angle (i.e., the surface is not rough) and all of the elements in the specimen must be considered by the algorithm.

Generally, random errors due to the finite number of X-rays in each peak are around ±1 at%. At best, with long counting times, ±0.1 at% may be possible. Random errors can be estimated from the counting statistics, although it is often better to analyse the same area a few times, or analyse similar areas in other parts of the specimen to get an idea of the spread in composition. Systematic errors, such as incorrect k-factors or poor absorption and fluorescence corrections due to non-uniform specimens or stray scattering cannot be estimated. Normally the systematic errors are greater than the random errors and they can only be compensated for by comparison with an area of similar but known composition.

X-ray mapping

X-ray mapping involves scanning an electron beam across an area of the specimen and measuring the composition at each point in order to create a map of the distribution of each element of interest across the specimen. This requires a microscope with STEM capability and a small focused probe.

Single X-ray spectra are typically taken over a period of 60 to 100s to obtain sufficient statistics. To produce an X-ray map with sufficient resolution in a reasonable time the dwell time for each point is typically no more than 0.1s. Thus high beam currents are required, often with some loss of spatial resolution, in order to decrease the noise. Even so, for older X-ray detectors it is normal for maps to have on a few X-ray counts per pixel for each element mapped.

Much higher count rates can be obtained with more recent large area X-ray detectors. These detectors are either close to the sample or multiple detectors arranged such that up to 1 steradian of solid angle can be used. These detectors are also silicon drift type detectors and can thus cope with X-ray signal rates of 10^5 counts per second or more allowing intense beams with currents up to 20nA.

With an aberration-corrected TEM and a relatively large beam current it is now possible to map the elemental distribution within the unit cells of many materials. An example of an atomic resolution X-ray map of $SrTiO_3$ imaged down [001] using an FEI Titan G2 TEM at 200kV is shown in figure 4.

3 Electron energy-loss spectroscopy (EELS)

EELS is an analysis technique, like EDX, where the energy of the electrons transmitted through the sample is measured in order to analyse samples in a TEM. Due to its use of transmitted electrons EELS can only be used only in a TEM [3].

Only the transmitted electrons close to the optic axis are used in EELS, typically up to scattering angles of only 10 or 20 mrad. This limits the amount of beam spreading as the electrons pass through the sample and thus potentially increases the spatial resolution of EEL spectra as compared to EDX spectra. The spatial resolution of EEL spectra is thus limited only by the size of the incident beam. EELS is very efficient – most of the electrons transmitted through the specimen enter the spectrometer. This should be compared with EDX, where X-rays emitted in all directions and only a small fraction are collected [4].

When an electron creates an X-ray, an Auger electron, a secondary electron or another process it loses energy. As a result the transmitted electrons contain information about all other interactions taking place in specimen. This makes EELS both very flexible, in that the spectra contain many different types of information, but also more complicated to understand and analyse than EDX.

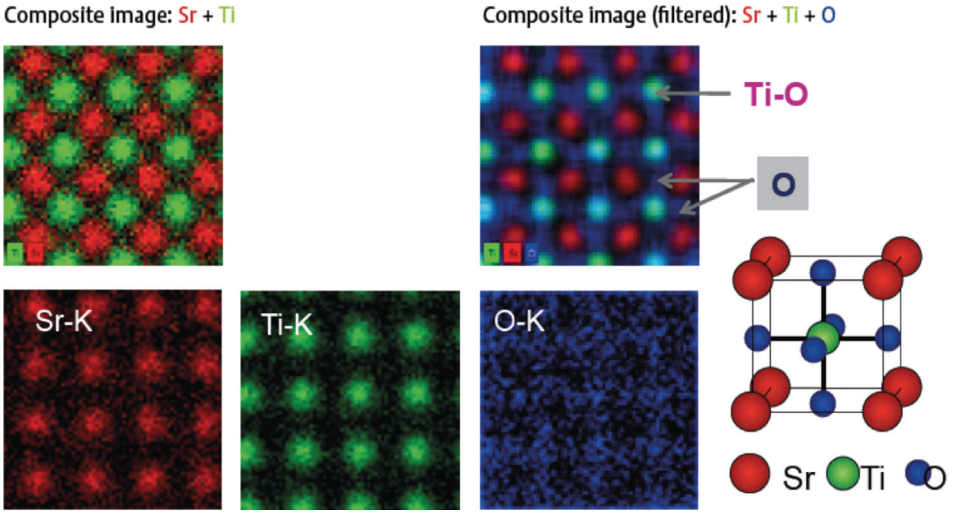


Fig. 4: *Atomic resolution EDX maps from a $SrTiO_3$ crystal imaged down the [001] axis [Bert Freitag, FEI].*

Although EELS can be used to measure elemental compositions, it can also give much more information. The shape of edges gives bonding and chemical information. Plasmon losses give compositional & electron density information. Very low losses give band structure information. And with an imaging filter it is possible to do zero loss imaging (for quantitative microscopy), plasmon loss imaging and elemental mapping (EFTEM).

Collecting EEL spectra

Electrons are transmitted through the specimen with a scattering angle of typically less than 1°. Therefore the spectrometer has to be after the specimen. It is typically located at the bottom of the microscope under the screen or part way down the column before the screen.

The energy-loss spectrometer (see figure 5) is a magnetic prism (or combination of prisms) which bends the electrons through (normally) 90°. Electrons with lower energy are bent through a greater angle giving a spectrum at the exit plane of the spectrometer. Slanted and curved ends of the magnetic prism focus the beam while extra optical elements are used to control the fine focus and remove aberrations. For a typical TEM the primary electron energy is 100 to 300kV and for optimal resolution the spectrometer energy resolution needs to be 0.5eV or better.

The spectrum is detected using a CCD camera, allowing the whole spectrum to be detected in parallel. To allow different energy ranges to be detected the spectrometer dispersion is controlled by a series of multipole lenses situated after the spectrometer.

An extension of a standard spectrometer is the imaging filter, where a further series of multipole lenses after the spectrum plane is used to re-form the original image on the TEM screen and remove distortions produced by the spectrometer magnet. In addition, a variable width slit at the spectrum plane can be used to select the electron energy range to be imaged.

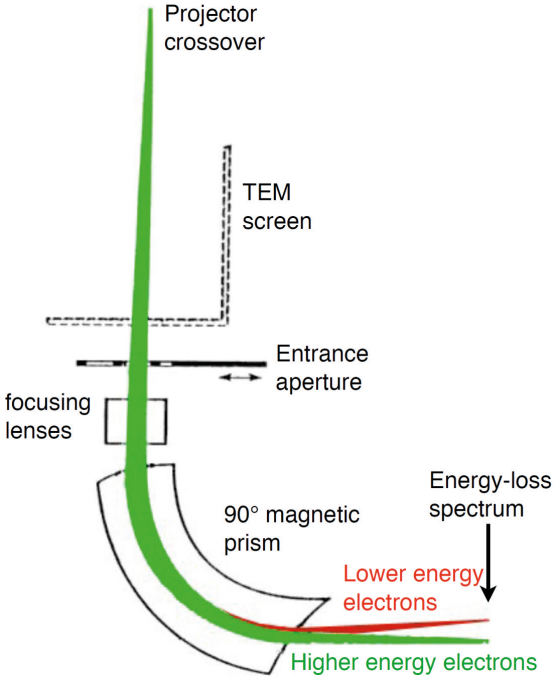


Fig. 5: *Diagram showing how a 90° prism electron energy-loss spectrometer works.*

Energy loss spectra

For most thin TEM specimens the majority of electrons pass through the specimen with no energy-loss causing interactions. Thus most of the electrons appear in the zero-loss peak, ie the peak corresponding to the incident electron energy (see figure 6). A small fraction lose some energy, with the fraction decreasing rapidly with increasing energy loss. This high dynamic range makes both collection and display of spectra difficult. In most cases the zero-loss peak is too intense for the CCD camera and as a result the energy range collected is adjusted so that it does not fall on the detector. Either a log scale or a number of gain changes is needed to display spectra.

There are 5 main regions to an EEL spectrum, as shown on figure 6 and these will be described in turn.

Zero-loss peak

The zero-loss peak contains both elastically scattered and phonon scattered electrons. Elastically scattered electrons pass through the specimen with no energy loss and include diffracted electrons. The energy width of the zero-peak is determined by energy spread of the electron gun. This ranges from ~2eV for a tungsten filament (thermal), ~1eV for a LaB_6 filament, ~0.7eV for a thermally assisted field emission gun (FEG), ~0.3eV for a cold field emission gun (cold FEG) and down to meV for monochromatic microscopes. The angular scattering is determined by diffraction in the sample, for example Si 111 (figure 7) has a scattering angle of 12mrad (0.7°).

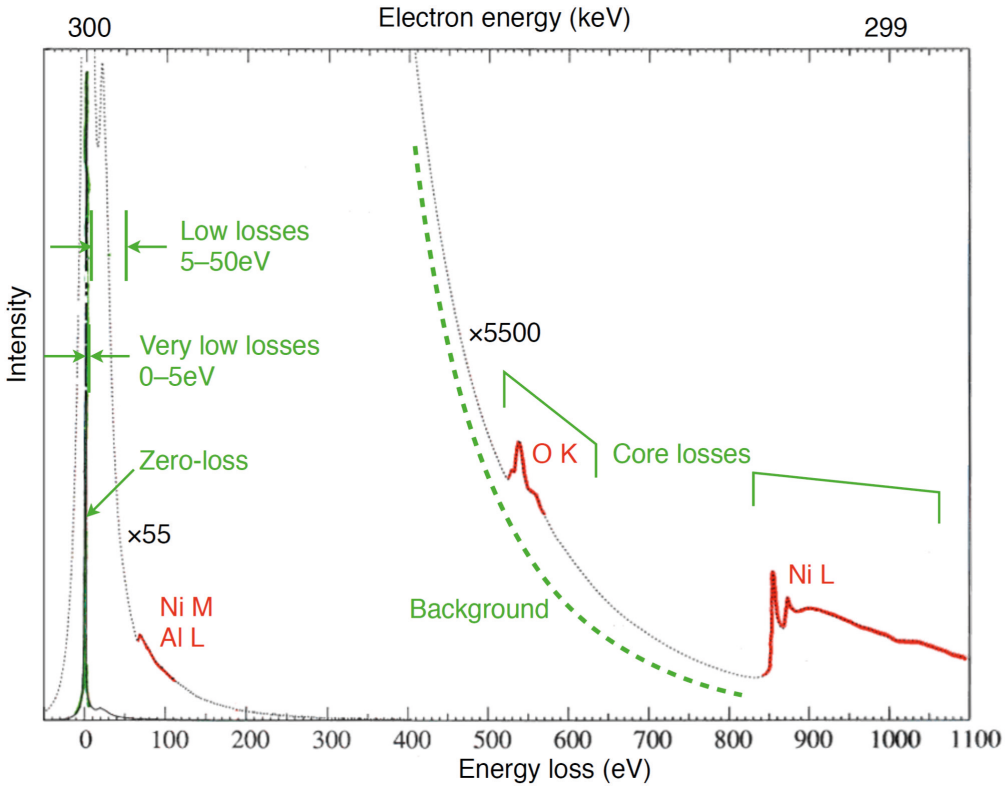


Fig. 6: *An example EEL spectrum from Ni_3Al showing the 5 main regions.*

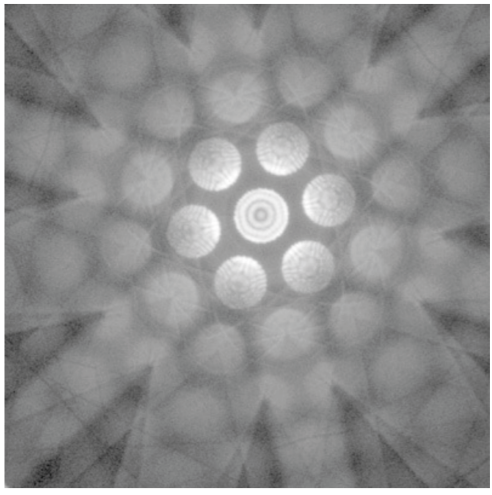


Fig. 7: *Zero-loss energy filtered convergent beam diffraction pattern from a Si crystal looking down [111]. The pattern was taken at 200kV and is shown on a log intensity scale to make the phonon scattering near the edges of the image visible.*

Phonon scattering is where an electron creates one or more phonons (thermal diffuse scattering). The typical phonon energy loss is ~kT ≈ 0.025eV, too small to measure except with the very latest monochromatic microscopes and thus lies within the zero loss peak, even though it is not actually zero energy loss. The angle of scattering is large, eg up to 50mrad (~3°). Phonon scattering is one source of Kikuchi lines in diffraction patterns (see figure 7).

Low loss region, 5 to 50eV

This region contains mostly plasmon scattering A plasmon is an oscillation of the conduction band electrons (like a phonon is an oscillation of the crystal lattice). The angular scattering for plasmon losses is a few mrad. Plasmon scattering is delocalised over a few nm.

Each compound has a characteristic plasmon energy, typically between 10 and 25eV. For example, the plasmon energy for Ta is 22eV while the plasmon energy for Ta oxide ranges from 22 to 27eV. Thus the plasmon energy can be used to distinguish the chemical state of a material, as for example in the resistive switching device whose EEL spectrum is shown in figure 8.

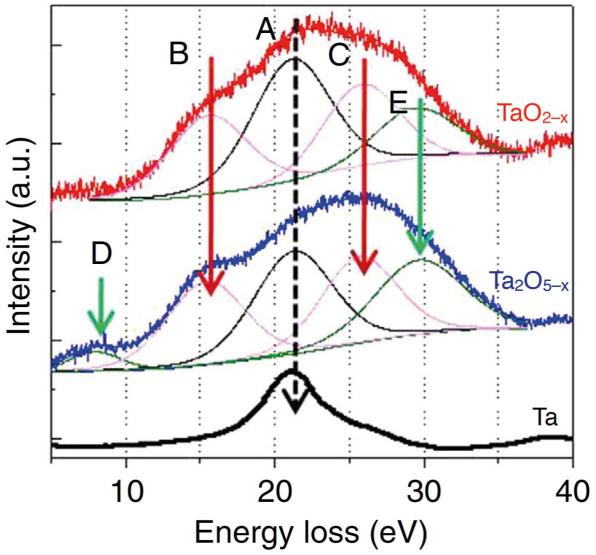


Fig. 8: *Low loss EEL spectra from Ta and Ta oxides from a resistive switching device [5].*

Very low loss region, 0 to 5eV

This region of the energy-loss spectrum has been explored relatively recently because a microscope with a monochromator or a cold field-emission gun microscope is required. In principle it should be possible to see semiconductor band gaps (such as Si at 1.1eV as shown in figure 9, or GaN at 3.5eV) plus the conduction band density of states at a spatial resolution of <1nm. However very low loss spectra are both difficult to obtain and difficult to interpret. In figure 9 the features between 2 and 4.5eV loss are not due to the band structure of Si but are due to Cherenkov radiation. Cherenkov radiation is light emitted when the velocity of the electron exceeds the velocity of light in the material. It can be minimised by using thinner samples. In addition, surface plasmon peaks are also present in this part of the spectrum with energies of up to 10eV. Surface plasmons can be minimised by using thicker samples.

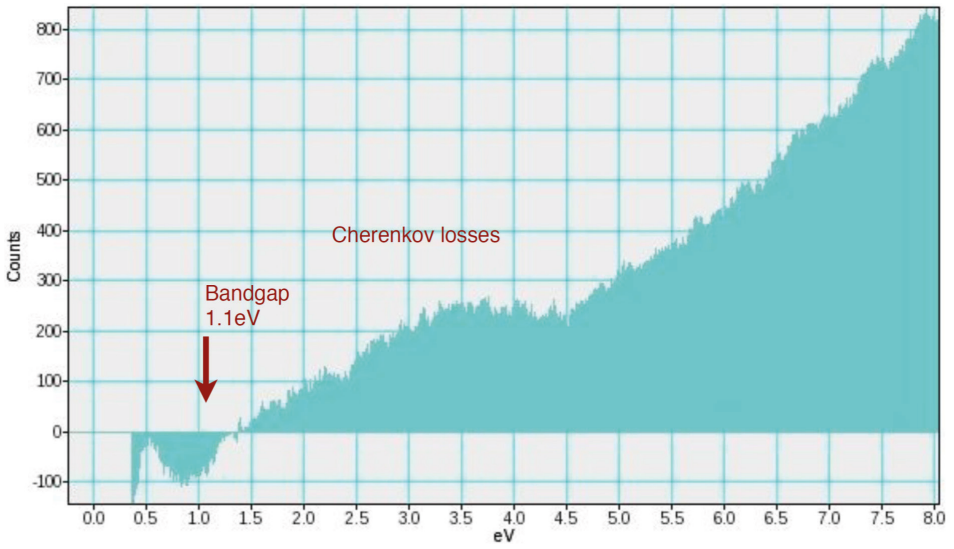


Fig. 9: *Very low loss EEL spectrum from Si taken with a monochromated microscope.*

Background, single electron excitations

Energy losses above 30eV are mainly by collision and ejection of single electrons from the valence band. These ejected electrons become the secondary electrons used for SEM imaging. The background intensity has a characteristic shape given by $I = AE^{-r}$, where I is the intensity, E is the energy loss, r is a constant, usually between 2 and 6 and A is another constant.

This formula enables the background to be subtracted from spectra. It is important for quantifying core losses. However it is only valid above about 50eV and for thin specimens.

Core losses

These are caused by the ejection of an inner shell core level electron from an atom to form a core hole with the incident electron losing energy in the process. The electron can be ejected to the conduction band (lowest empty state) or to higher energies. As a result an "edge" rather than a peak is seen. Each atomic shell (K, L, M, etc) has its own EELS edge or edges. Later an electron from a higher shell fills this hole causing X-ray or Auger electron emission. The energy of the emitted X-ray is always less than corresponding EELS edge energy. Electron energy loss spectra are similar to X-ray absorption spectra, but can be obtained from a much smaller volume. An example of an EEL spectrum from copper showing the Cu L and Cu M edges is shown in figure 10.

Quantifying EEL spectra

Quantification of EEL spectra to obtain elemental compositions requires the areas of the core-loss edges to be measured. But quantifying energy loss spectra is more difficult than for X-ray spectra, mainly because the EELS edges extend over a wide range of energy, unlike the sharp X-ray peaks making the edge areas difficult to measure. Also, it is only possible to fit a background above the edge energy.

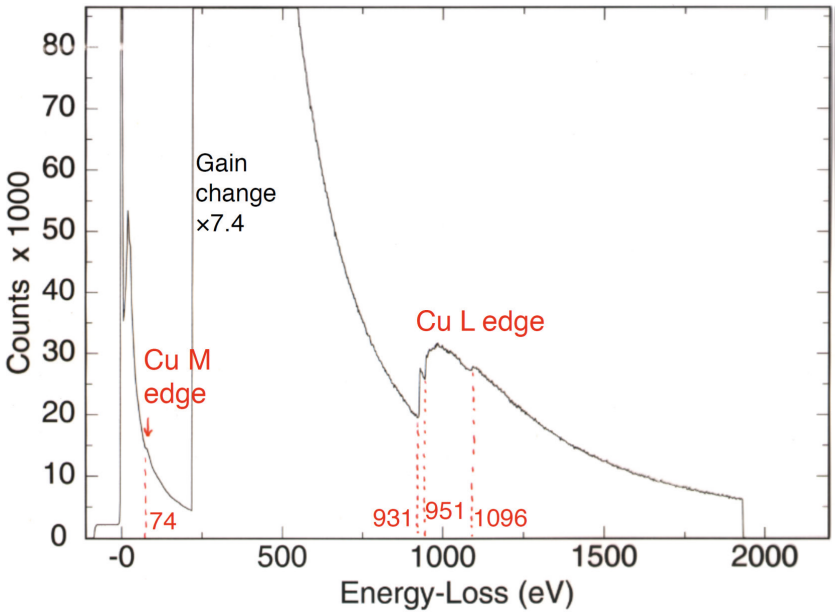


Fig. 10: *Core loss EEL spectrum from Cu.*

To quantify energy-loss spectra the specimen must be thin or else multiple scattering will obscure the edges. Ideally, the plasmon peak should be less than 1/10 of the zero loss peak. First the background must be fitted using a background fitting window above the edge of interest. Then the background is subtracted before the area of the edge within a window can be found (see figure 11). For each edge of interest the ionisation cross-section, σ, must be calculated and integrated over the same window.

Compositions can be calculated using a ratio technique similar to that used for EDX, $C_A/C_B = (I_A/\sigma_A)/(I_B/\sigma_B)$, where C_A is the concentration of element A, I_A is the counts under the edge and σ_A is the ionisation cross-section.

The cross-sections are difficult to calculate accurately. To calculate the cross-section, we need to know the collection angle (ie the objective aperture size or EELS entrance aperture size) and the microscope voltage. Cross-sections are larger for lower Z elements (but the background also larger). Typically the accuracy is around 10%, worse than for EDX, due to inaccurate cross-sections and the difficulty of estimating edge areas accurately.

Electron energy-loss near edge structure (ELNES)

EELS edges have some fine structure which is a function of the local atomic arrangement around the edge element. Energy loss near edge structure (ELNES) is the structure within about 50eV of the edge onset. Extended energy loss fine structure (EXELFS) is the structure beyond 50eV after the edge onset and depends on the arrangement of the surrounding atoms. An example of a K edge from Al showing both ELNES and EXELFS is shown in figure 12. Only ELNES will be considered here.

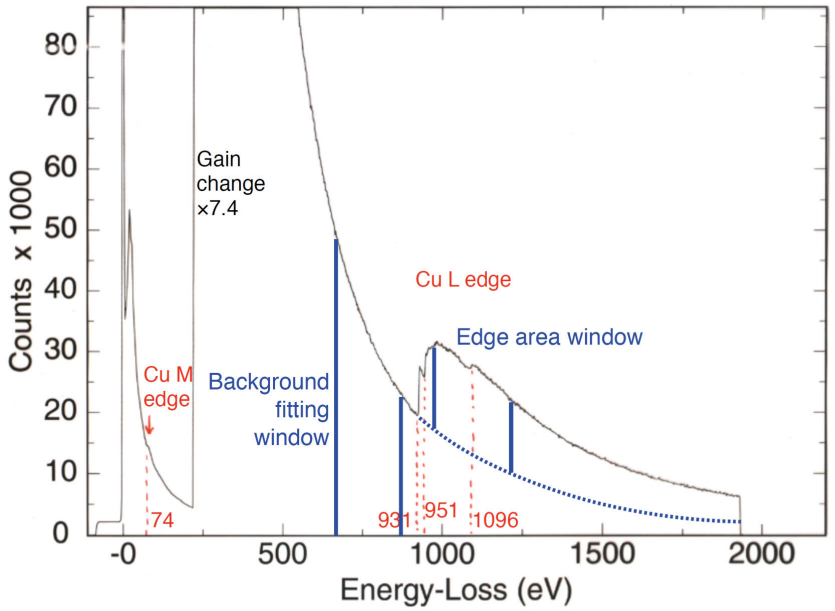


Fig. 11: *Cu EEL spectrum showing background fitting and edge area windows.*

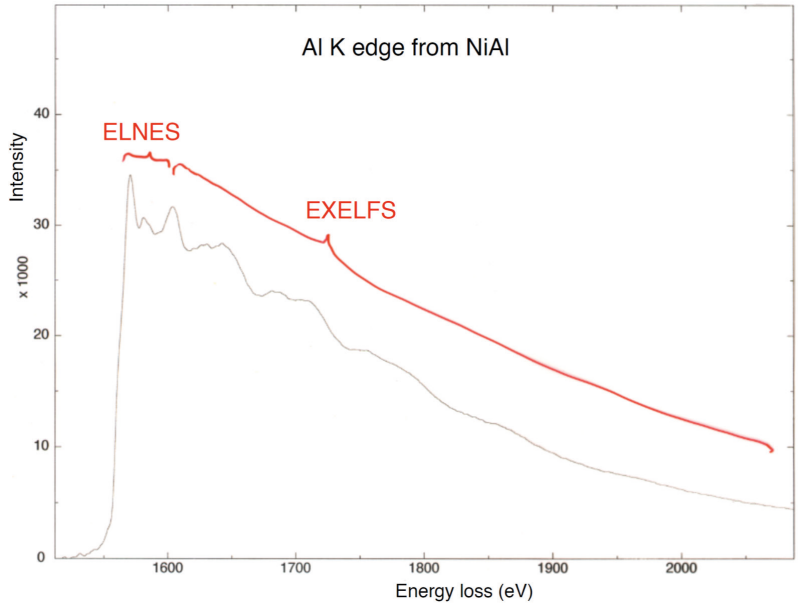


Fig. 12: *Al K EELS edge from NiAl showing the ELNES and EXELFS regions.*

In EELS an electron is excited from a core level to the first empty state of the atom, the conduction band. The shape of edge depends on both the initial state (the core level) and the final state (the conduction band). For K edges (eg Al) there is only 1 core level so the edge shape depends only on the conduction band density of states. For L edges there are 3 initial states. However the ELNES structure is not the same for all the three initial states because the quantum mechanical selection rules allow different transitions for each initial state. Therefore ELNES only shows a partial density of states for the conduction band.

For many materials different bonding states of the same atom can give very different ELNES structures. An example is carbon and spectra from a number of different types of carbon are shown in figure 13.

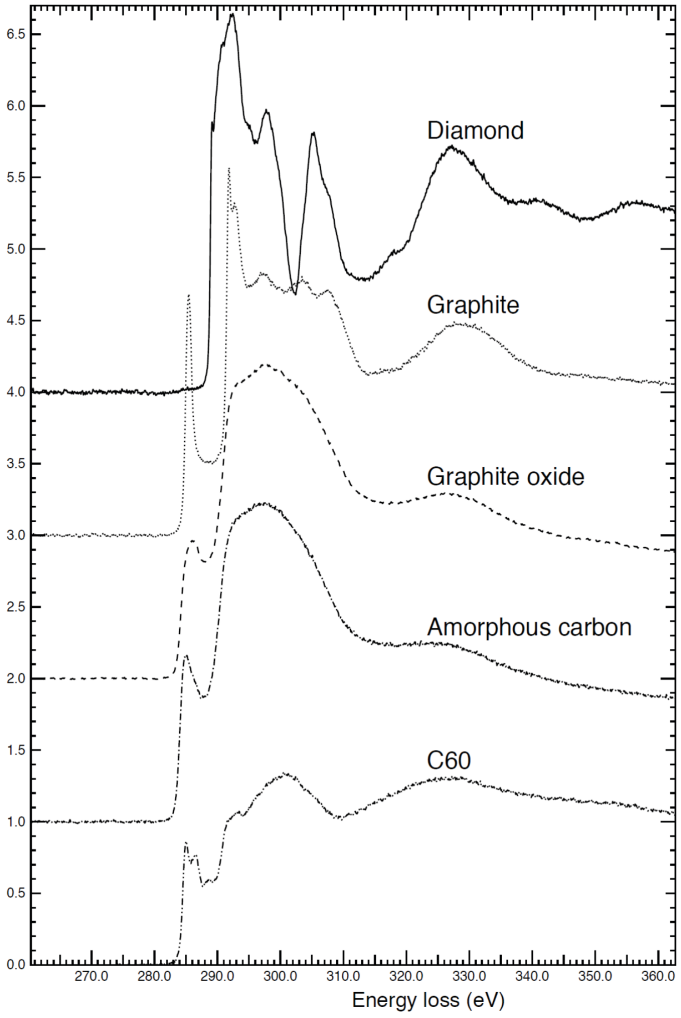


Fig. 13: *Carbon K edge EEL spectra from diamond, graphite, amorphous carbon, graphite oxide and C60.*

Energy-filtered electron microscopy (EFTEM)

Energy filtering is an extension of energy loss spectroscopy. In energy filtering, we are able to collect an image from an area of the specimen at a given energy loss. This can be done either in STEM or TEM mode. In STEM mode a small focused probe is used and an energy loss spectrum is collected through a slit at the desired energy loss. The probe is scanned to create an image. This is often called "spectroscopic imaging". It uses a standard energy-loss spectrometer but needs a TEM with STEM mode.

Alternatively in TEM mode extra lenses after the energy-selecting slit can be used to allow the original image to be reformed. The extra lenses form part of an imaging filter. This method effectively collects an image in parallel at each energy loss. A number of different uses for EFTEM will now be considered.

Zero-loss filtered images

In zero loss filtering the energy-selecting slit is positioned so as to collect only electrons that have lost no energy, ie those in the zero loss peak. This means that all the inelastically scattered electrons are prevented from reaching the image allowing thicker regions to be examined (loss electrons suffer chromatic aberration & thus blur image). Zero-loss filtering also makes quantitative TEM possible, since most image simulations assume only elastic scattering. An example is the improvement in visibility of convergent beam diffraction patterns from thick specimens shown in figure 14.

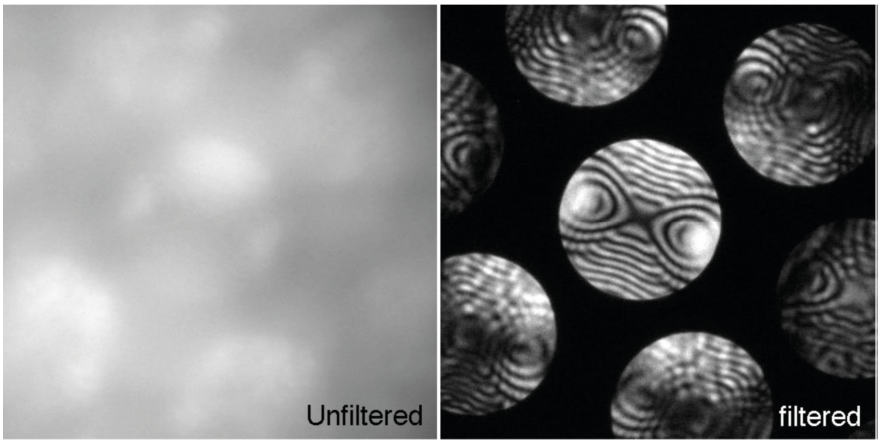


Fig. 14: *Comparison of unfiltered (contains electrons of all energies) and zero-loss (contains only zero-loss electrons) energy filtered convergent beam diffraction patterns from a silicon crystal with the beam along [110].*

Plasmon loss images

The plasmon peaks are characteristic of the different compounds in the sample. Thus plasmon images can distinguish different compounds. Figure 15 shows a film of SiO_2 that has been irradiated in the centre to form Si particles. These show up as the bright dots in the Si plasmon image while the SiO_2 plasmon image shows only the surrounding undamaged SiO_2. Although the resolution of plasmon loss images is lowered by delocalisation, it is good enough to see individual Si particles of a few nm.

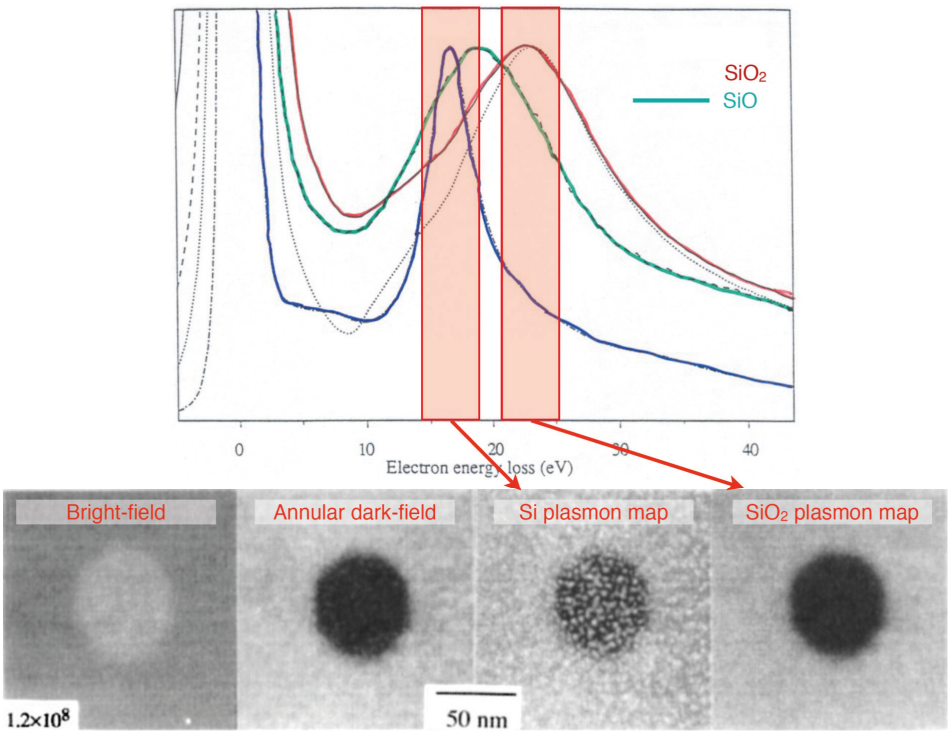


Fig. 15: *Bright-field, annular dark-field STEM image and Si and SiO_2 plasmon loss images from a SiO_2 film that has been irradiated with electrons in the centre.*

Core loss images

The aim of core loss mapping is to collect a map of a particular core loss edge and thus also an elemental map. Core loss mapping is difficult because of the need to remove the background from under the core loss edge, which may be much higher in intensity than the core edge itself. There are various methods of calculating core loss maps.

For the jump ratio method, one image just before the edge (pre-edge image) and one image on the edge (post-edge image) are collected. The jump ratio image is the ratio of these images. A jump ratio image only gives an approximate idea of the elemental concentration. Its advantage is that thickness variations are removed (at least approximately).

For the three-window method 2 pre-edge images and one post-edge image are collected. The 2 pre-edge images are used to estimate the background under the post-edge image using the formula $I = AE^{-r}$. This background can then be subtracted from the post edge image. The method is equivalent to the normal background subtracting method of finding EELS edge areas from energy loss spectra and can thus be quantified. The 3-window image is the sum of the edge area over the thickness of the specimen. Hence to get true elemental maps any specimen thickness variations must be compensated for. Figure 16 shows an example of core-loss elemental maps from a TiN/HfO_2-based resistive switching structure. The O K map shows localised oxygen deficiency in the HfO_2 layer.

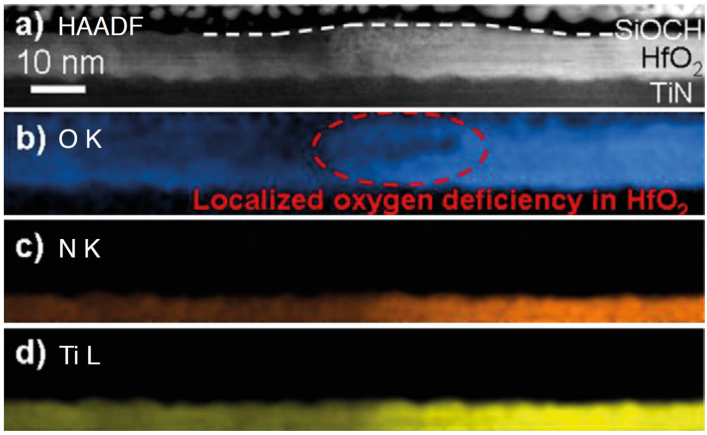


Fig. 16: *Core loss elemental maps from a TiN/HfO_2-based resistive switching structure. (a) HAADF STEM image, (b) O K, (c) N K and (d) Ti L elemental maps [6].*

Image spectroscopy

In image spectroscopy an entire EEL spectrum is collected for each point in an image. This can be done either in TEM mode where many images are collected both before and after the edge of interest or in STEM mode where a complete spectrum is collected for every point as the beam is scanned across the specimen. Spectrum images collected in TEM mode have good spatial resolution (typically 2k by 2k pixels) but poor energy resolution and can suffer from alignment problems if the sample drifts. Image spectra collected in STEM mode have good energy resolution but the scanned area is limited by the speed that the spectra can be read out, so usually are not more than about 200 by 200 pixels. Drift in STEM mode causes the spectrum image to be spatially distorted.

Image spectroscopy thus gives a big improvement over the 3-window method in determining the background although it takes longer and requires a larger dose of electrons. As a result the signal to noise ratio is better and the resulting elemental maps are more quantitative than for either the jump-ratio or 3 window methods.

As an example figure 17 shows atomic resolution maps of the V, La and Ti concentrations from a $LaVO_3/SrTiO_3$ multilayer derived from a STEM spectrum image collected using a Nion SuperSTEM operated at 100kV.

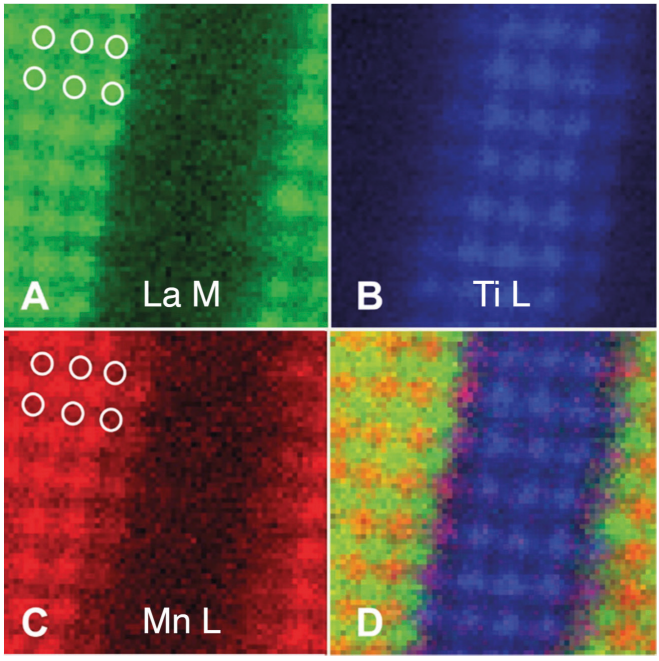


Fig. 17: *A-C, Atomic resolution La, Ti and Mn elemental maps from a $La_{0.7}Sr_{0.3}MnO_3$/ $SrTiO_3$ multilayer derived from a STEM spectrum image. D, False colour image obtained by combining the three maps.*

References

[1] J Goldstein, DE Newbury, DC Joy, CE Lyman, P Echlin, E Lifshin, LC Sawyer and JR Michael, Scanning Electron Microscopy and X-ray Microanalysis (Springer, 2003).

[2] PJ Goodhew, J Humphreys and R Beanland, Chapter 6, Electron Microscopy and Analysis (CRC Press, 2000).

[3] RF Egerton, Electron energy loss spectroscopy in the electron microscope (Springer, 2011)

[4] DB Williams and C Barry Carter, Transmission Electron Microscopy: A Textbook for Materials Science (Springer, 2009).

[5] G-S Park, YB Kim, SY Park, XS Li, S Heo, M-J Lee, M Chang, JH Kwon, M Kim, U-I Chung, R Dittmann, R Waser and K Kim, Nature Comm. 4 (2013) 2382.

[6] P Calka, E Martinez, V Delaye, D Lafond, G Audoit, D Mariolle, N Chevalier, H Grampeix, C Cagli, V Jousseaume and C Guedj, Nanotechnology 24 (2013) 085706.

C 5 Photoelectron Spectroscopy

L. Plucinski
Peter Grünberg Institut
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Fundamental Aspects of Photoemission **4**
2.1 Three-Step Model of Photoemission 7
2.2 One-Step Model of Photoemission 12
2.3 Refinement of the One-Electron Model 13

3 Techniques **14**
3.1 Electron Spectrometers 14
3.2 Spin Analysis 16
3.3 Ambient Pressure PES 17

4 Selected Examples **18**
4.1 Electronic and Chemical States 18
4.2 Spin Effects in Photoemission 24
4.3 Electronic Correlations 31
4.4 Kinkology 33
4.5 High-Energy Photoemission (HAXPES) 35
4.6 Interfacial sensitivity 39

5 Conclusions **41**

Lecture Notes of the 47$^{\text{th}}$ IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

Photoelectron spectroscopy has matured into an extremely versatile and powerful analysis technique. It permits access to a very wide variety of materials and their electronic structure, ranging from complex bulk structures down to free atoms. Consequently, there is an enormous wealth of results and interesting examples on different systems available, which are well worth being discussed. For good reasons, however, this lecture must focus on a few essential basics, novel aspects and a very personal selection of examples. For an in-depth study of photoemission spectroscopy and phenomena, the reader is referred to a number of excellent textbooks and review articles covering this field [1, 2, 3, 4, 5, 6].

Moreover, this lecture leaves out recent developments in spectromicroscopy, where photoemission microscope (PEEM) can be used both in the real and in the reciprocal space (angular) mode [7]. This subject is covered in this issue in the lecture **C6** by C. M. Schneider.

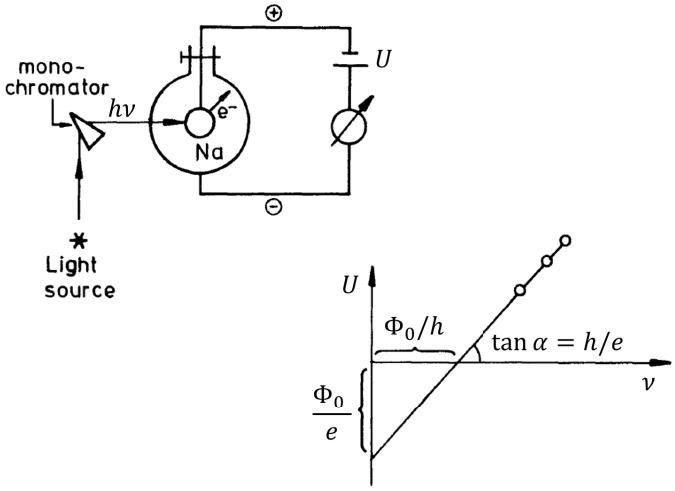


Fig. 1: *Early photoemission experiment. Monochromatic photons of the energy $h\nu$ excite alkali (K, Na) sample and the retarding voltage U is applied until the photoemission current disappears. Dependence of U on the frequency ν is linear and the slope yields the Planck's constant h, while the offset yields the work function Φ_0. Figure adapted from [1].*

The method of photoelectron emission spectroscopy goes back to the photoelectric effect, which was discovered by H. Hertz in 1887 [8] and refers to the phenomenon of electrons being ejected from a metal when illuminated by electromagnetic radiation. The explanation of the photoelectric effect by A. Einstein in 1905 [9], along with Compton's work on inelastic X-ray scattering (published in 1923) [10], were essential discoveries which confirmed corpuscular nature of light within the frame of the discussion on wave-particle dualism. In his famous paper, Einstein extended Planck's quantum hypothesis by postulating that quantization was not a property of the emission mechanism, but rather an intrinsic property of the electromagnetic field. Using this hypothesis, Einstein was able to explain why the maximum kinetic energy E_{kin} of the emitted electrons varies with the frequency ν of the incident radiation as

$$h\nu = \Phi_0 + E_{kin} = \Phi_0 + \frac{1}{2} m_e v^2 \equiv eU \tag{1}$$

where h is Planck's constant, Φ_0 is a characteristic energy and called the work function, m_e, e and v denote electron mass, charge and velocity, respectively. This is exactly the result expected if photons are quantized with energies $h\nu$, and earned Einstein the 1922 Nobel prize in physics. In the early photoemission experiments one basically determined the total photoelectron current $I_p(U)$ on a counter electrode as a function of a retarding voltage U. In this way, the maximum kinetic energy of the photoelectrons was determined from the condition $I_p(U) \Rightarrow 0$ (Fig. 1). From today's perspective this approach corresponds to an angle-integrated photoemission experiment [1].

The development of photoelectron spectroscopy started at the end of the 1950's with K. Siegbahn, who studied the energy levels of core electrons in atoms using excitation with X-rays [11]. Since the exact binding energy position of the core level depends on the chemical environment of the atom from which the photoelectron is emitted, Siegbahn coined the name Electron Spectroscopy for Chemical Analysis (ESCA) for this spectroscopic technique. He was awarded the physics Nobel prize in 1981 for his contributions to high-resolution electron spectroscopy.

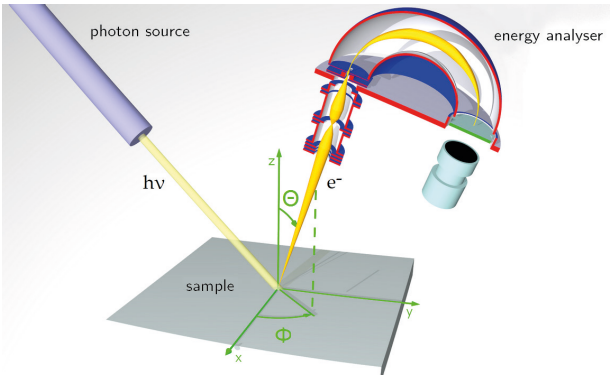


Fig. 2: *Schematic picture of a modern photoemission experiment.*

A typical angle-resolved photoemission experiment as of today is sketched in Fig. 2. The photon source is typically a synchrotron radiation facility, which provides light over a broad range of photon energies $h\nu$ from the ultraviolet up to hard X-rays [12]. This light beam can be finely focused down to about 100 μm in diameter, its direction of incidence onto the sample and its degree and orientation of polarization (linear, circular) can be precisely controlled. The electron energy analyzer is equipped with an electron optical entrance lens system, which defines the angular range of the photoelectrons analyzed. Therefore, the emission direction denoted by the angles (θ, ϕ) is a free experimental parameter. On the one hand, for solid state experiments, mostly single-crystalline samples are investigated under ultrahigh vacuum conditions ($p < 10^{-8}$mbar). On the other hand, studies in catalysis require almost atmospheric pressure a the sample. This can be nowadays enabled by special designs of the entrance lens system involving differential pumping stages [13].

2 Fundamental Aspects of Photoemission

The valence electronic structure of a solid can – in principle – be calculated by solving a Schrödinger (nonrelativistic approximation) or Dirac equation (relativistic interactions included) for the respective lattice structure. For most metallic systems with weak electronic correlations, i.e. being close to the limit of a homogeneous electron gas, a quite successful description has been achieved within the framework of density functional theory (DFT) using the local density (LDA) or more general local spin density approximation (LSDA) for the exchange-correlation potential [14]. In order to discuss salient features of the photoemission process, we will first adopt this effective single-particle picture, although we must be aware of the fact that it does not capture electronic correlations properly. Introduction to theoretical methods of calculating the electronic structure is given in the lecture **A2** by S. Blügel and G. Bihlmayer

One important result of the theoretical treatment is that the quantum mechanical wave function, describing an electronic state in the solid, depends on symmetries of the Hamiltonian (e.g. lattice symmetries, inversion symmetry, time reversal symmetry, etc.) and must also include electron spin. Within a single particle picture the Hamiltonian can be written as

$$\Big(\Big[\tfrac{1}{2m}\left(\mathbf{p}-\tfrac{e}{c}\mathbf{A}\right)^2+eV(\mathbf{r})\Big]+i\tfrac{e\hbar}{4m^2c^2}\mathbf{E}\cdot\mathbf{p}-\tfrac{e^2\hbar}{2mc}\boldsymbol{\sigma}\cdot\mathbf{B}- -\tfrac{e\hbar}{4m^2c^2}\boldsymbol{\sigma}\cdot(\mathbf{E}\times\mathbf{p})\pm eV_{exc}^{\uparrow\downarrow}(\mathbf{r})\Big)\,\phi=E_{nls}(\mathbf{k})\phi. \tag{2}$$

with ϕ and $E_{nls}(\mathbf{k})$ denoting the single electron wave function and energy eigenvalue, respectively. The terms in square brackets in Eq. 2 represent the Hamiltonian of a system subjected to an electromagnetic field (vector potential $\mathbf{A}$). This part contains all crystalline symmetries through the potential $V(\mathbf{r})$. The Darwin term ($\sim \mathbf{E}\cdot\mathbf{p}$) may be understood as a relativistic correction to the electron energy. The fourth term contains the interaction of the spins – described by the Pauli spin matrices $\boldsymbol{\sigma}$ – with an external magnetic field $\mathbf{B}$. The last two terms contain the spin-dependent interactions through spin-orbit coupling and the exchange-correlation potential $V_{exc}^{\uparrow\downarrow}(\mathbf{r})$. The latter is responsible for the formation of spontaneously ordered magnetic states in solids. All spin-dependent terms in the Hamiltonian tend to reduce the symmetry of the system in one way or the other, leading to the splitting and hybridization of degenerate states. A full set of energy eigenvalues obtained from the DFT treatment forms a band structure $E_n(\mathbf{k})$ of the solid with the band index n and the electron wave vector $\mathbf{k}$. The wave functions are Bloch functions and are further classified by the orbital momentum quantum number ℓ and the spin quantum number s. Formally, the respective states may be written as $|n,\ell,\mathbf{k},s\rangle$ with their energy eigenstates $E_{nls}(\mathbf{k})$.

In addition to this valence electronic states comprising delocalized electrons, the full electronic structure of a solid also contains atomic-like localized core levels at higher binding energies $E_B \gtrsim 30$ eV (the binding energy is referred to the Fermi energy, i.e. $E_B = 0$ at E_F). In the ground state, the core states are completely occupied, whereas the valence states are occupied up to the Fermi energy E_F in the case of metals. In semiconductors and insulators the Fermi level lies in a band gap and the intrinsic bulk states are occupied only up to the valence band edge in the undoped case. Moreover, in semiconductors band bending and surface photovoltage phenomena occur, which complicate the intepretation of photoemission results from semiconductors[15]. These effects will not be treated in this contribution.

Interaction of such electronic structure with photons leads to excitation of electrons from occupied into unoccupied states in the electronic structure. If these empty states are located above

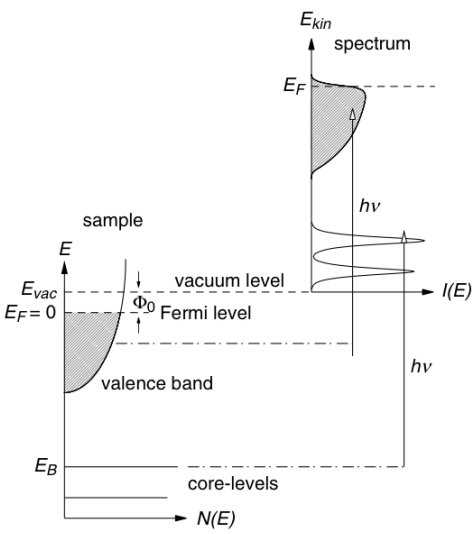


Fig. 3: *Principle of the photoemission process. The electrons excited into states above the vacuum level form a photoelectron spectrum reflecting a broad valence electronic distribution (shaded area) and sharp emission lines from the core levels. From [3].*

the vacuum level E_{vac} of the solid, photoelectrons can leave the crystal and can be measured by an electron spectrometer, yielding characteristic signatures of the valence electronic states and core levels (Fig. 3). Fermi's Golden Rule describes quantum mechanically the transition probability between two electronic levels $|i\rangle$ and $|f\rangle$ with binding energies E_i and E_f, respectively:

$$P_{i\to f} = \frac{2\pi}{\hbar} \left|\langle f| \mathcal{O} |i\rangle\right|^2 \delta(E_f - E_i - h\nu) \tag{3}$$

In the simplest approach, the two levels $|i\rangle$ and $|f\rangle$ may be taken from the ground state electronic structure of the solid – which neglects the role of electronic correlations in the excitation process, as we will see below. The most important quantity in Eq. (3) is the transition matrix element

$$M_{fi} = \langle f| \mathcal{O} |i\rangle \tag{4}$$

which depends on the symmetries of the electronic wave functions and the photonic operator $\mathcal{O}$, whereas the delta function $\delta(E_f - E_i - h\nu)$ ensures energy conservation in the excitation process. For low photon flux densities the operator $\mathcal{O}$ can be treated within linear response theory and takes the form

$$M_{fi} = \frac{-e}{mc} \langle f| \mathbf{A}(\mathbf{r}) \cdot \mathbf{p} |i\rangle \tag{5}$$

with $\mathbf{A}(\mathbf{r})$ the vector potential of the electromagnetic field and $\mathbf{p}$ the momentum operator. It is usually assumed that the wavelength of the electromagnetic field is large compared to interatomic distances, i.e. $\mathbf{A}(\mathbf{r})$ varies only marginally in the spatial region contributing to the transition matrix element [1]. This view is commonly known as *dipole approximation* and simplifies the transition matrix element to

[1] This assumption should be revisited, if we go to photoexcitation with hard X-rays.

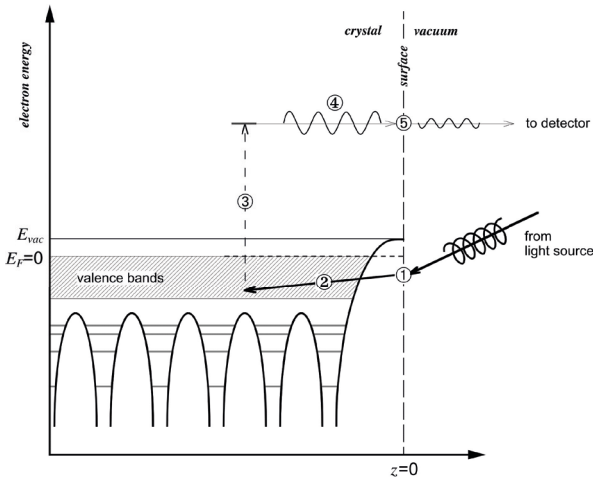


Fig. 4: *Schematic representation of the three step model. The numbers denote: (1) refraction of the electromagnetic wave at the surface, (2) penetration of the photon into the solid, (3) photoexcitation, (4) propagation of the photoelectron to the surface, and (5) diffraction of the electron wave at the surface.*

$$M_{fi} = \frac{-ie}{\hbar c} A_0 (E_f - E_i) \langle \psi_f | \, \mathbf{e} \cdot \mathbf{r} \, | \psi_i \rangle \tag{6}$$

with the complex amplitude of the vector potential A_0, its polarization vector **e**, and the wavefunctions of the final and initial states ψ_f and ψ_i, respectively. This form of the transition matrix element is extremely valuable, as the quantity $\langle \psi_f | \, \mathbf{e} \cdot \mathbf{r} \, | \psi_i \rangle$ can be evaluated for selected symmetries of the wave functions and yields dipole selection rules, which are very useful for a qualitative interpretation of photoemission spectra. This can be most easily seen for atomic levels, the wavefunctions of which can be expressed in terms of a radial part and a part containing spherical harmonics $Y_{l,m}$. As the operator $\mathbf{e} \cdot \mathbf{r}$ can also be represented in terms of spherical harmonics (e.g. $Y_{1,0}$ for linearly polarized light, or $Y_{1,\pm m}$ for circularly polarized light), the matrix element $\langle \psi_f | \, \mathbf{e} \cdot \mathbf{r} \, | \psi_i \rangle$ can be fully calculated by evaluating products of spherical harmonics. The particular mathematics of spherical harmonics allows the matrix element to be nonzero only for particular relations between l_f, l_i, m_f, m_i, which is the basis of the selection rules. Although being only strictly valid for atomic systems, this approach has also been successfully extended to approximately describe the behavior of electronic states at high symmetry points in solids. A more general treatment of dipole selection rules in solids has been developed on the basis of group theory [1].

It is useful to recall that the photoexcitation is only part of the entire photoemission process. Once the electron has been excited into the upper level – which takes place on a timescale of 10^{-15} s – we call it a photoelectron. However, this highly energetic or "hot" photoelectron must find a way to leave the crystal, which is only possible if the excitation occurs into states above the vacuum level E_{vac}. The proper quantum mechanical treatment of the photoemission process is the so-called *one-step model*, which – at least in principle – permits a quantitative interpretation of photoemission spectra. A more intuitive access to the underlying physics is provided by the simpler *three-step model* which we will discuss in the following.

2.1 Three-Step Model of Photoemission

This model separates the photoemission of a single[2] electron into subsequent processes dealing with (i) the *photoexcitation*, (ii) the *transport* of the hot electron to the surface, and (iii) the *transmission* of the electron through the surface into the vacuum (Fig. 4). On a quantum mechanical level these steps have to be connected in a suitable way in order to allow the electronic wave function to propagate from one step to the next.

2.1.1 Photoexcitation

We have already mentioned above that the central aspect in the photoexcitation step concerns dipole selection rules. These rules predict allowed electronic transitions based on the symmetry of the electronic wave functions involved. In the atomic picture, these selection rules take the form:

$$\Delta L = \pm 1 \tag{7}$$

$$\Delta m_L = 0, \pm 1 \tag{8}$$

The photon carries an amount of angular momentum of $|L| = 1$ with it's polarization state being determined by $m_J = 0$ (linear polarization) and $m_J = \pm 1$ (right- and left-hand circular polarization, respectively). Linearly polarized light can be represented as a superposition of right- and left-hand circularly polarized waves. To illustrate the action of these selection rules we take the example of the excitation from an atomic $2p$ level. According to Eq. (7) we will find two types of allowed transitions, which may contribute to the photoemission spectrum

$$p \rightarrow \begin{cases} d \text{ for } \Delta L = +1 \\ s \text{ for } \Delta L = -1 \end{cases} \tag{9}$$

For the evaluation of Eq. (8) it is useful to consider that atomic states are usually subject to spin-orbit coupling, which leads to a characteristic splitting of the atomic levels and leaves only the total angular momentum $J = L + S$ as a good quantum number. Consequently, our p-level splits into a $p_{3/2}$ and a $p_{1/2}$ state and for linearly polarized light, we will have allowed transitions of the type

$$\begin{array}{lcl} p_{3/2} & \rightarrow & d_{3/2} \\ p_{1/2} & \rightarrow & s_{1/2} \\ p_{-1/2} & \rightarrow & s_{-1/2} \\ p_{-3/2} & \rightarrow & d_{-3/2} \end{array} \tag{10}$$

whereas for circularly polarized light we have

[2] This single-electron picture is convenient, because in many cases it allows a qualitative interpretation of photoemission spectra on the grounds of band structure calculations within the framework of density functional theory. It neglects, however, electronic correlations in the electronic structure which can be significant in certain materials or material classes.

$$\begin{array}{lcl} p_{3/2} & \rightarrow & d_{5/2} \\ p_{1/2} & \rightarrow & d_{3/2} \\ p_{-1/2} & \rightarrow & s_{1/2} \\ p_{-3/2} & \rightarrow & s_{-1/2} \end{array} \text{ for } (\Delta m_L = +1) \quad \text{and} \quad \begin{array}{lcl} p_{3/2} & \rightarrow & s_{1/2} \\ p_{1/2} & \rightarrow & s_{-1/2} \\ p_{-1/2} & \rightarrow & d_{-3/2} \\ p_{-3/2} & \rightarrow & d_{-5/2} \end{array} \text{ for } (\Delta m_L = -1) \quad (11)$$

Note that the dipole operator of the light acts only on the orbital part of the electronic wave function, i.e. on the spatial symmetries, but it cannot interact with the electron spin S directly. However, because spin-orbit coupling ties the spin to specific orbitals, a selective excitation can yield spin polarized photoelectrons even from nonmagnetic materials. This phenomenon is called optical spin-orientation [16] and is also the basis of all magneto-dichroic effects observed in photoabsorption and photoemission [17]. In order to see how this works let us have a closer look at the $p \rightarrow s$ transitions described by Eq. (11) (selection rules for crystalline symmetries see Appendix). The states $s_{1/2} \equiv |\uparrow\rangle$ and $s_{-1/2} \equiv |\downarrow\rangle$ may be regarded as pure spin states. For positive light circularity we have transitions starting at $p_{-3/2}$ and $p_{-1/2}$. Usually, the probabilities for the two transitions - which can be simply calculated from the Clebsch-Gordan coefficients [18] - differ by a factor of 3, i.e. the amount of photoelectrons excited into the $s_{-1/2}$ is 3 times greater than into the $s_{1/2}$ state. If we assume that we have a nonmagnetic situation, the $s_{-1/2}$ and $s_{1/2}$ states will be energetically degenerate and a summation over the two photocurrent contributions will yield a spin polarization of the photoelectrons of $P = -50\%$. The same treatment for negative light circularity yields $P = 50\%$, i.e. a reversal of the circularity also reverses the sign of the photoelectron spin polarization. This optical spin-orientation effect in the $p \rightarrow s$ transitions is particularly exploited in spin-polarized GaAs photocathodes [19], but it can also be observed as a general photoemission phenomenon in basically all materials.

The quantity measured in the photoemission experiment is a photocurrent $I(h\nu)$, which is composed by transitions between all possible initial (i) and final states (f)

$$I(h\nu) \sim \sum_{i,f} |\langle f| \mathcal{O} |i\rangle|^2 \, \delta(\epsilon_f - \epsilon_i - h\nu) \quad (12)$$

Note that in this single particle picture the photoemission spectrum is represented by a series of sharp lines, which is not what is observed in the experiment. The reason of this discrepancy is the many-electron nature of the photoemission process, which will be discussed in sect. 2.3.

Once the photoelectron has been excited into the upper state it will propagate through the solid with a kinetic energy of $E_{kin} = h\nu - E_B$ according to the energy conservation. The propagation direction is determined by the electron momentum $\hbar\mathbf{k}$, i.e. the electron wave vector $\mathbf{k}$ inside the crystal, which in turn is determined by the wave vector conservation law $\mathbf{k}_f = \mathbf{k}_i + \mathbf{k}_{h\nu} \pm \mathbf{G}$. The relation between initial and final state wave vectors $\mathbf{k}_f$, $\mathbf{k}_i$ and the photon momentum $\mathbf{k}_{h\nu}$ is given modulo a reciprocal lattice vector $\mathbf{G}$. In most cases, we can therefore confine our considerations on the photoemission spectra to the first Brillouin zone (reduced zone scheme). For low photon energies well below 1000 eV the photon momentum can be safely neglected and particularly with respect to electronic band structures one then may assume *vertical* transitions, directly connecting the initial and final state. For higher photon energies, however, the photon momentum may become a significant quantity.

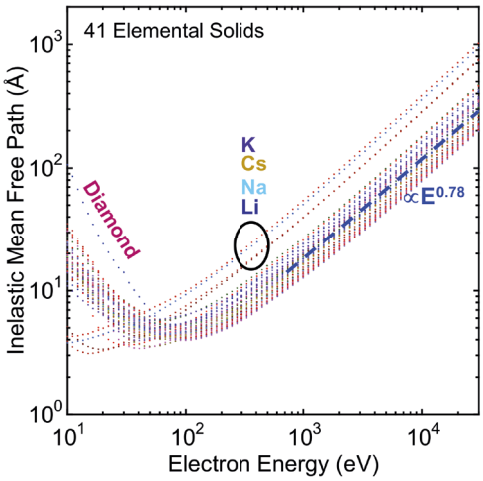


Fig. 5: *IMFP values for 41 elements, calculated using the TPP-2M formula: Li, Be, three forms of carbon (graphite, diamond, glassy C), Na, Mg, Al, Si, K, Sc, Ti, V, Cr, Fe, Co, Ni, Cu, Ge, Y, Nb, Mo, Ru, Rh, Pd, Ag, In, Sn, Cs, Gd, Tb, Dy, Hf, Ta, W, Re, Os, Ir, Pt, Au, and Bi. Five "outlier" elements (diamond and the alkali metals) are included to illustrate the influence of the electronic structure characteristics. The dashed straight line for higher energies represents a variation as* $\lambda_{in} \sim E_{kin}^{0.78}$*, and is a reasonable first approximation to the variation for all of the elements shown. From [20].*

2.1.2 Propagation

Inelastic mean free path – The propagation of the hot electron is described in the second step of the three-step model. Due to the strong Coulomb interaction in a solid the hot electron will suffer very efficient elastic and inelastic scattering processes, which affect both the energy and angular distribution of the photoemission spectrum observed outside the crystal. The main mechanisms are scattering due to electron-electron interactions and scattering on defects. In particular, the inelastic scattering processes lead to a relaxation of the photoelectron towards the Fermi level. The effect of the inelastic scattering can be described by exponential damping of the photoelectron intensity along the path l

$$I(l) = I_0 \exp(-\frac{l}{\lambda_{in}}) \tag{13}$$

with the quantity λ_{in} being the *inelastic mean free path* (often abbreviated as IMFP). The concept of the inelastic mean free path is essential in describing the finite information depth in a photoemission experiment and λ_{in} is the average distance between two subsequent inelastic scattering events. This quantity is generally believed to follow a very similar behavior in different materials, which leads us to the well-known *universal curve* of the energy dependence of λ_{in}. A closer look, however, reveals that the curve is universal only with respect to the general shape and depends on the electronic structure of the element or material in question (Fig. 5). Common to the IMFP curves is a minimum of λ_{in} of only a few Ångstroms at kinetic energies of ~100 eV and an increase towards both lower and higher energies. Based on this one can understand the high surface sensitivity of photoelectron spectroscopy at intermediate energies, which is both a virtue and a limitation. It can be – at least partly – overcome by exciting the photoelectrons with hard X-ray photons, i.e. photon energies of 6 - 10 keV (see Chapter 4.5). From Fig. 5 we see that at a kinetic energy of 10 keV the IMFP increases up to 10 nm. This energy-dependent variation of the information depth forms the basis for hard X-ray photoemission.

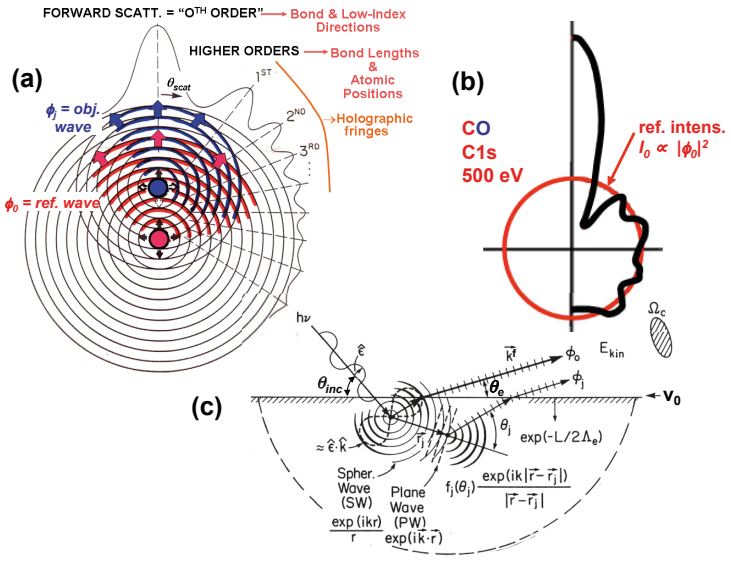


Fig. 6: *Illustration of various aspects of photoelectron diffraction. (a) Simple diffraction features expected in emission from one atom in a diatomic system. (b) An accurately calculated diffraction pattern for C 1s emission from CO at a kinetic energy of 500 eV. Note the strong forward scattering peak, and other interference peaks or fringes extending from near the forward scattering direction to the backward scattering direction. (c) The basic theoretical measures required to describe photoelectron diffraction. From Ref. [21], with calculations via Ref. [22].*

Photoelectron diffraction – While moving through the crystal the photoelectron can also be elastically scattered, giving rise to photoelectron diffraction (PD). This phenomenon is often also referred to as X-ray photoelectron diffraction (XPD) due to the higher excitation energies that are often used. In XPD a core-level photoelectron scatters from the atoms neighboring the emission site, so as to produce an angular anisotropy in the outgoing photoelectron intensity [21]. The qualitative effects expected for the simple case of emission from the bottom atom in the diatomic molecule are shown in Fig. 6(a), and a quantitative calculation for emission from the C 1s subshell in an isolated CO molecule at 500 eV kinetic energy is shown in Fig. 6(b). Electron-atom elastic scattering is typically peaked in the forward direction, with this effect becoming stronger (that is, having a stronger and narrower forward peak) as energy increases [21]. For the CO case in Fig. 6(b), the intensity in the forward direction is in fact enhanced relative to that expected without scattering (I_0 in the figure) by about three times. Thus, one expects in XPD curves both a forward scattering peak (also referred to as forward focussing) along the interatomic direction, as well as higher-order diffraction interference effects that one can also consider to be holographic fringes. Back scattering is weaker as energy increases, but Fig. 6(b) shows that, even at 500 eV, there are still interference fringes in the backward direction.

Such XPD effects are very useful to determine the local atomic arrangement around an emitter atom. The XPD signals can be interpreted and modelled using the quantities shown in Fig.

6(c). The polarization $\hat{\varepsilon}$ of the light influences the directionality of the initial photoelectron wave, and for emission from an s-subshell, the outgoing unscattered wave φ_0 has an amplitude proportional to $\hat{\varepsilon} \cdot \hat{k}$, where $\hat{k}$ is a unit vector in the direction of the photoelectron wave vector, and the photoelectron deBroglie wavelength will be given by

$$\lambda_e = h/\left|\mathbf{p}\right| = 2\pi/\left|\mathbf{k}\right|, \quad \text{in convenient units}: \ \lambda_e[\text{Å}] = \sqrt{150.5/E_{kin}[\text{eV}]} \tag{14}$$

Thus, an electron with 150 eV kinectic energy has a wavelength of about 1 Å, and a 1500 eV electron of about 0.3 Å, and these numbers are comparable to atomic dimensions. The outgoing photoelectron will elastically scatter from neighboring atoms j to produce wave components φ_j, and this process can be in first approximation described by plane-wave scattering, or more accurately by spherical-wave scattering. This scattering can be incorporated into a scattering factor f_j, which is furthermore found to be strongly peaked in the forward direction for energies above about 500 eV, as noted previously. The photoelectron wave components will also be inelastically attenuated as they traverse some total path length l in getting to the surface, with their *amplitudes* decaying as $exp(-l/(2\lambda_{in}))$. Finally, they will be refracted at the inner potential barrier V_0. Summing up all wave components (unscattered and scattered) and squaring then yields the diffraction pattern. Electrons can also be multiply scattered from several atoms in sequence, and in many cases accurate calculations of the resulting photoelectron diffraction patterns require including these effects, especially if scatterers are lined up between the emitter and the detection direction, as along low-index directions in multilayer emission from a single crystal. Various programs are now available for calculating XPD patterns, with one web-based version being particularly accessible [22].

2.1.3 Transmission

In the final step of the three-step model the photoelectron has arrived at the surface and will leave the crystal. For this purpose, however, it has to pass through the surface potential barrier which matches the periodic potential inside the crystal to the vacuum outside. From simple quantum mechanics we know that an electron wave passing across a potential step Φ will be elastically scattered and diffracted, i.e. it will change its trajectory. In reality the surface potential barrier

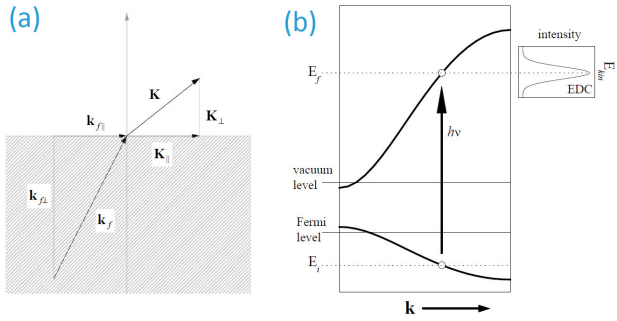


Fig. 7: *(a) The component of the wave vector parallel to the surface is conserved upon transmission of the electron through the surface, $\boldsymbol{k}_{f||} = \boldsymbol{K}_{||}$. (b) If photon momentum is neglected, transitions between the initial and final bands of the crystal are vertical in the reduced zone scheme.*

is not a step function, but smoothly varies as $\Phi(z)$. The details of the scattering process depend on the shape of $\Phi(z)$.
Let us use the following nomenclature for the considered wave vectors which is illustrated in Fig. 7(a): $\mathbf{k}_i$ describes the initial state of the electron inside the solid before the excitation, $\mathbf{k}_f$ describes the final state of the electron inside the solid after the excitation ("hot electron"), but before the transmission through the surface, and $\mathbf{K}$ describes the electron outside the solid. The diffraction of the photoelectron at the surface potential is the reason that the wavevectors of the electrons inside the crystal $\mathbf{k}_f$ and outside in the vacuum $\mathbf{K}$ are not conserved. A conservation law exists only for the component parallel to the surface plane, $\mathbf{k}_{f||} = \mathbf{K}_{||}$.
This conservation of the parallel momentum makes it natural to consider components parallel and perpendicular to the surface separately. In order to relate the perpendicular component of the photoelectron wave vector $K_\perp$ to the electronic states inside the solid, more sophisticated methods for the band mapping have to be used [1]. The simplest one assumes the final electronic states to be described by a nearly-free electron parabola

$$E_{kin} + V_0 = \frac{\hbar^2}{2m}\mathbf{k}_f{}^2 = \frac{\hbar^2}{2m}(\mathbf{k}_i + \mathbf{G})^2 \tag{15}$$

with an electron mass m and with the inner potential V_0, the empirical parameter which describes the energy loss during the transmission through the surface potential barrier (here we neglected photon momentum $\mathbf{k}_{h\nu}$).
As illustrated in Fig. 7(b) the initial and final states are connected by vertical transitions in the band structure, $\mathbf{k}_i$ for the certain emission angle can be determined from E_{kin} under the assumption of a certain V_0. Assuming that only the shortest $\mathbf{G}$ vector which allows photoemission is involved ($\mathbf{G} = G_\perp$, a so-called first Mahan cone) one arrives to the simple equations which relate the electrons of the certain kinetic energy E_{kin}, emitted at the angle θ with respect to the surface normal, to their wavevector $\mathbf{k}_i$

$$k_{i||} = \sqrt{\frac{2m}{\hbar^2}E_{kin}}\sin\theta \tag{16}$$

$$k_{i\perp} = \sqrt{\frac{2m}{\hbar^2}(E_{kin}\cos^2\theta + V_0)} \tag{17}$$

This procedure fails, however, if the final state band structure deviates considerably from the nearly-free electron picture, which is the case particularly at hybridization regions. In this case, it is more reasonable to use final states from a band structure calculation for comparison. However, all these analyses will only yield a qualitative interpretation of the photoemission spectra. For a quantitative interpretation, a full photoemission calculation within a one-step model is needed.

2.2 One-Step Model of Photoemission

A considerable drawback of the three-step model is its limitation to a qualitative description of the photoelectron spectra. For a more quantitative description, first of all, the transition probabilities for all electronic excitations must be calculated on the basis of a realistic band structure for a semi-infinite system, for example, derived from density functional theory. The formalism must also take into account surface states or transitions into evanescent final states.

Secondly, we must properly consider the multiple scattering events which the hot electron suffers in the final state. These are caused by the strong electron-electron interaction. This situation is more adequately described by the so-called one-step model, which considers the excitation and subsequent transport in a common framework. In particular, the multiple scattering of the electron waves in the surface-near region and in the surface potential is treated in analogy to formalisms developed for low energy electron diffraction (LEED). In LEED an electron wave enters the crystal and subsequently undergoes a multiple scattering process. If we invert the order on the time scale, we retrieve our final state in the photoemission process, with the excited electron propagating towards the surface. This state is therefore also called a *time-reversed LEED state* [23]. In terms of kinetic energy there is a gradual transition from the LEED to the photoelectron diffraction regime. PED effects can therefore be included into one-step photoemission theories.
On the basis of one-step photoemission theories one arrives at a quantitative description of the photoelectron spectra. This may even include effects due to spin-orbit coupling and the electron spin, in which case a relativistic Dirac-type formalism is involved [24]. In this way it becomes possible to calculate magnetic dichroism and spin polarization spectra.

2.3 Refinement of the One-Electron Model

2.3.1 Electronic correlations

In the above discussions we have always implicitly assumed that the electronic system under investigation can be modeled within an effective single-electron picture. This has the advantage that the features appearing in the photoemission spectra may be directly related to specific interband transitions in the electronic structure, involving band states or core levels. Single-particle picture may fail to capture the essential physics of a system, because it may underestimate the correlations in a many-electron system. This is true for the entire family of so-called highly-correlated systems, which includes transition metal oxides and other materials exhibiting phenomena such as high-temperature superconductivity, colossal magnetoresistance or multiferroicity. Such systems are usually described by theoretical approaches beyond simple LDA, for example, LDA+U or dynamical mean field theory (DMFT). The interpretation of photoemission results from such systems is more involved and must take into account the influence of the electron-electron interactions in all steps of the photoemission process.

2.3.2 Spectral shape of photoemission lines

There is, however, a second way through which electronic interactions enter the photoemission experiment. According to Fermi's Golden Rule (Eq. 3) the energy conserving δ-function implies all photoemission signatures to be infinitely sharp lines. This should hold particularly for core level photoemission lines. Inspection of the schematic picture in Fig. 3 already reveals that the photoemission line will have a finite width. This is due to the multielectron character of the photoemission process itself.
Whenever a photoelectron is excited to the upper level, it leaves behind a hole in the lower level. Strictly speaking the photoemission process converts an N-electron system into an $(N-1)$-electron system, if the photoelectron has left the crystal, before the hole has been filled again. The photoelectron and the hole interact with each other through Coulomb interaction, which may lead to a renormalization of the binding energies, the appearance of spectral satellites, and a finite linewidth of the spectral line. This has two profound consequences. First, the terms

ground and excited state become a different meaning, because they rather refer to an N and $(N-1)$ electronic system, respectively. Second, photoelectron spectroscopy always measures an excited state of matter rather than the electronic ground state. In a somewhat larger picture this is a nice illustration of one of the paradigms in quantum physics, according to which a measurement always affects and alters the system measured. Fortunately, modern condensed matter theory is able nowadays to handle many-electron systems both in the ground and excited state and can therefore provide a full photoemission calculation.

The role of electronic interactions in the photoexcitation spectrum is often taken into account within a Green function formalism [2]. The Green function $G(\mathbf{k},\epsilon)$ describes the behavior of a quasiparticle, which is "dressed" by the electronic correlations and the electron-hole interaction. Their influence is globally expressed by means of the complex self-energy Σ. Without going through the details of the formalism, from the Green function one can finally calculate the *spectral density function* $A(\mathbf{k},\epsilon)$, which may be compared to experimental results

$$A(\mathbf{k},\epsilon) = -\frac{1}{\pi}\mathrm{Im}G(\mathbf{k},\epsilon) = -\frac{1}{\pi}\frac{\mathrm{Im}\Sigma(\mathbf{k},\epsilon)}{[\epsilon-\epsilon_k-\mathrm{Re}\Sigma(\mathbf{k},\epsilon)]^2+[\mathrm{Im}\Sigma(\mathbf{k},\epsilon)]^2}. \tag{18}$$

A closer inspection of Eq. 18 reveals that the real part of Σ introduces a renormalization of the energy ϵ_k of the spectral feature, whereas the imaginary part of Σ describes the finite lifetime of the quasiparticle state, resulting in a finite spectral width. As we will see below, the self-energy Σ can be conveniently employed to include further interactions, such as electron-phonon and electron-magnon coupling.

The spectral density function replaces the delta function in Eq. 12. The total photocurrent is again determined by summing up over all dipole-allowed optical transitions between the many-electron states Φ_f and Φ_i weighted by the spectral density. We then arrive at the following description of the photocurrent [25]

$$\begin{aligned} I(h\nu) &\sim \sum_{f,i} |\langle\Phi_f|\,\mathcal{O}\,|\Phi_i\rangle|^2\, A_{ii}(\epsilon_f - h\nu) \\ &= \frac{1}{\pi}\sum_{f,i} |\langle\Phi_f|\,\mathcal{O}\,|\Phi_i\rangle|^2 \frac{|\mathrm{Im}\Sigma(\epsilon_i)|}{[\epsilon_f-\epsilon_i-h\nu-\mathrm{Re}\Sigma(\epsilon_i)]^2+[\mathrm{Im}\Sigma(\epsilon_i)]^2}. \end{aligned} \tag{19}$$

This is the basis for modern photoemission calculations, which attempt a quantitative interpretation of the experimental data.

3 Techniques

3.1 Electron Spectrometers

During the last two decades, there has been a considerable improvement in photoelectron spectrometer technology, mainly driven by the continuous quest for improved spectral resolution. By now, commercially available energy analyzers may be able to achieve an energy resolution below 1 meV [26]. The most common type of photoelectron spectrometers nowadays are display-type energy filters which are able to efficiently acquire intensity distributions over a certain range of angles and energies within a single measurement. An example for such a hemispherical display spectrometer is given in Fig. 8. The photoelectrons moving away from the sample surface are accepted by a lens system, which defines the angular spread, i.e. the $k_\parallel$

value transmitted. The hemispherical capacitor employed to disperse the electrons according to their kinetic energy is usually operated at a fixed pass energy E_{pass} in order to keep the energy resolution constant throughout a spectrum. The slit between the lens and the hemispheres separates the angular and energy information. The lens system therefore also has the task to accelerate/decelerate the electrons to the pass energy. After being dispersed in the electrostatic field a part of the electrons leaves the analyzer through a second aperture towards the areal electron detector. This usually comprises a combination of a multichannel plate (MCP) and position-sensitive read-out. The MCP consists of an array of narrow channels each being typical several 10 μm in diameter. In each channel a photoelectron is amplified by a factor of $10^4 - 10^6$. This signal is then transported into the position-sensitive readout. The read-out may be a phosphor screen observed by a CCD camera system which sorts and counts the events into a data file in a computer. Alternatively, there are resistive anode type detectors, which directly output voltage pulses to a multichannel analyzer.

The specific design shown schematically in Fig. 8 features another anode arrangement called a delayline detector (DLD) [27, 28] to make room for a second detector measuring the spin polarization of the photoelectrons.

For specific purposes a wide variety of specialized electron spectrometers have been developed over the years. Most of them employ the electrostatic dispersion principle or a time-of-flight approach, in which the kinetic energy of the electrons is converted into a transit time along a defined trajectory. DLD detectors are in particular suitable for time-of-flight spectrometers, because of their intrinsic time resolution.

One of the challenges is to increase the angular acceptance of the analyzer in order to be able to capture a larger part of the photoelectron angular distribution in front of the sample. A very

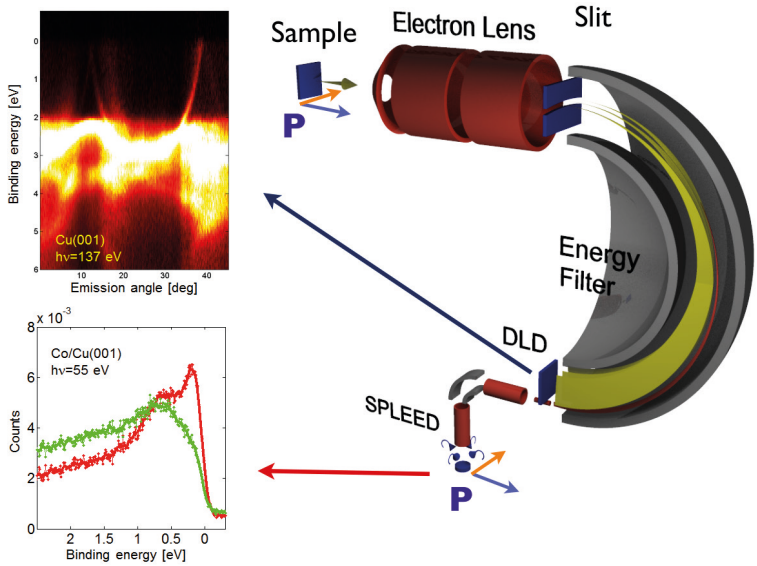


Fig. 8: *Hemispherical photoelectron spectrometer with two-dimensional delayline detector (DLD) and SPLEED spin polarization analyzer operated at the synchrotron laboratory DELTA in Dortmund by the institute PGI-6 [28].*

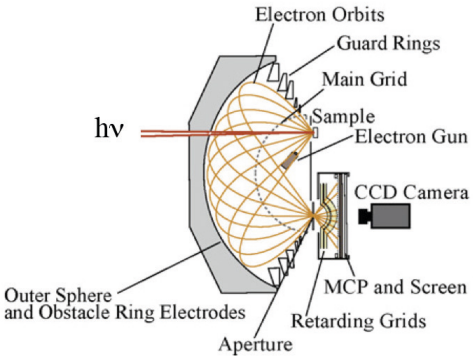


Fig. 9: *Cross sectional scheme of the DIANA spherical capacitor spectrometer. The angular distribution of the photoelectrons is imaged onto a two-dimensional detector and captured by a CCD camera. From [29].*

interesting design in this respect is the display-type spherical analyzer DIANA (Fig. 9) [29]. The electron trajectories emerging from the sample surface even at large emission angles are guided with high angular fidelity into the detector. The spectrometer is capable of mapping almost the entire half-space in front of the sample. This is particulary useful, for example, for photoelectron diffraction studies.

3.2 Spin Analysis

For the sake of completeness photoemission experiment should be capable of analyzing the photocurrent with respect to all quantum numbers $|n, \ell, \mathbf{k}, s\rangle$. This includes the electron spin, which carries important information about spin-dependent excitation and scattering processes in the solid. It can be shown that for an ensemble of electrons, the quantity spin can be expressed by a vector in real space, the spin polarization **P**, the direction of which is defined by a quantization axis in the solid or the entire experiment [30]. Spin-dependent effects arise either through spin-orbit coupling or exchange interaction, the latter being a characteristic quantity in magnetic systems.

Several types of spin polarization analyzers have been developed over the years. Their common principle of operation is based on the spin-dependent scattering of the photoelectrons off a target. The spin-dependence in the scattering process comes about by the same spin-dependent interactions mentioned above. In a simple picture, these interactions define a spin quantization axis and cause electrons with spin-up and spin-down to scatter with different probability into a direction perpendicular to this quantization axis. A counting detector placed in this direction will thus count different rates of scattered electrons for incident spin-up or spin-down photoelectrons, for example, $I^{\uparrow}(E)$ and $I^{\downarrow}(E)$. By subsequently orienting the spin-sensitive axis of the detector along the x, y, and z-axis, we can determine all three components of the spin polarization vector $\mathbf{P}(E)$.

There is only one spin polarization analyzer so far which is based on the exchange interaction. It involves low energy scattering ($E_s \approx 6$ eV) at a single-domain Fe(001) surface. Detectors exploiting spin-orbit coupling either involve high-energy Mott scattering at the atomic potential (several 10 keV up to 100 keV) or low energy scattering at the periodic potential of a solid (typically 100eV). Since the strength of the spin-orbit coupling increases with the atomic number *Z*, all spin-orbit scattering targets comprise heavy atoms, such as Au, W, or U.

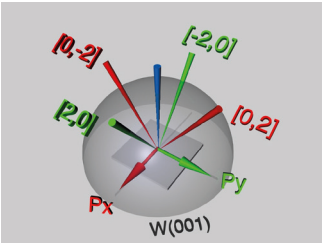


Fig. 10: *Sketch of the SPLEED spin detection principle using a W(001) crystal. The spin polarization components P_x (red) and P_y (green) are measured simultaneously.*

In order to see how a spin detector is interfaced with the electron spectrometer, we choose the SPLEED detector as an example (Fig. 10). In this detector one effectively performs a **s**pin-**p**olarized **l**ow-**e**nergy **e**lectron **d**iffraction experiment [31]. The incoming electrons hit a W(001) surface at normal incidence with a scattering energy of about 104 eV. The diffracted beams create a four-fold symmetric LEED diffraction pattern above the surface. The $\{20\}$ diffraction beams are of particular importance, because they provide the highest spin sensitivity at these scattering conditions. Because of symmetry reasons the SPLEED detector is sensitive to two orthogonal components of the spin polarization vector. The components P_x and P_y are determined from the intensity of the LEED reflexes by

$$P_x = \frac{1}{S} \cdot \frac{I_{[0,2]} - I_{[0,-2]}}{I_{[0,2]} + I_{[0,-2]}} \qquad (20)$$
$$P_y = \frac{1}{S} \cdot \frac{I_{[2,0]} - I_{[-2,0]}}{I_{[2,0]} + I_{[-2,0]}}$$

with the spin sensitivity S. This procedure is repeated for every data point of the spectrum and yields a spin polarization spectrum $P_{x,y}(\mathbf{k}, E)$, which can be used to calculate the spin-up and spin-down contributions of one vector component to the photoemission spectrum according to

$$I^{\uparrow}(E) = \frac{I_0}{2}(1 + P(E)) \quad \text{and} \quad I^{\downarrow}(E) = \frac{I_0}{2}(1 - P(E)) \qquad (21)$$

with the spin-averaged total intensity I_0.

3.3 Ambient Pressure PES

One of the limitations of standard X-ray photoemission (XPS) is the need for ultra-high vacuum (UHV) experimental conditions which makes it impossible to investigate surfaces under atmospheric pressures. One reason for UHV is that in many experiments one is interested in atomically clean surfaces, due to the surface sensitivity of XPS. On the other hand, understanding elemental composition and chemical specificity of surfaces under nearly atmospheric pressures is of primary interest in the fields of energy generation and heterogeneous catalysis. Ambient-pressure X-ray photoelectron spectroscopy (AP-XPS), pioneered by Kai Siegbahn's group at Uppsala University [32], has been developed in order to allow for photoemission measurements under nearly atmospheric conditions. Electrons are strongly scattered by gas molecules, and this scattering is the main challenge in AP-XPS. For 100 eV electrons in 1 mbar water vapor the inelastic mean free path is approx. 1 mm [13]. This is much shorter than the

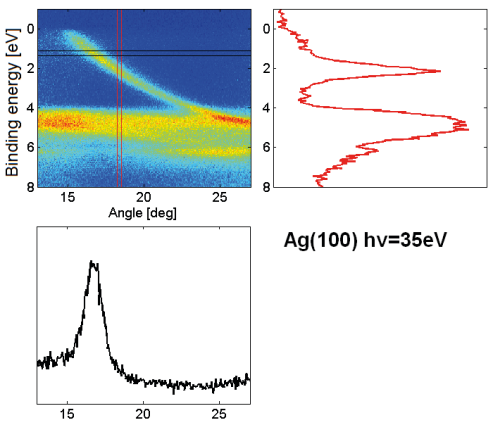


Fig. 11: *Example of a two-dimensional $E(\theta)$ distribution recorded from an Ag(100) single crystal surface. The color code ranges from blue (no intensity) through yellow (medium intensity) to red (high intensity). Vertical and horizontal cuts through this distribution yield energy distribution curves (EDC) and momentum distribution curves (MDC), respectively. The broken lines bound areas of 5 lines on the detector which have been added up to the MCD (top) and EDC (right). Inset: Experimental geometry with the red triangle indicating the angular spread measured.*

distance between the sample and the entrance of the lens of the hemispherical analyzer, such as the one shown in Fig. 8, which is typically in order of 3-4 cm.

AP-XPS systems use differential pumping between the sample environment, which is called "*in situ* cell", and the photoelectron spectrometer and the photon source (laboratory-based X-ray source or the synchrotron beam). Such pumping schemes typically use small apertures which separate two or three differentially pumped subsections. One can either mount the apertures in front of the lens of the existing spectrometer, which has the advantage of minimal modifications, or integrate the pumping stages and apertures into the electrostatic lens. Currently AP-XPS systems can operate at pressures up to 130 mbar.

Comprehensive review of the AP-XPS technique has recently been given by Starr *et al.* [13]. AP-XPS can also be used to investigate the liquid-solid interfaces [33].

4 Selected Examples

In the following, we will discuss several examples illustrating different applications of photoemission spectroscopy covering band states and core levels.

4.1 Electronic and Chemical States

4.1.1 Valence state photoemission

The first example illustrates the mapping of the valence electronic states in a noble metal. A hemispherical display-type spectrometer like the one shown in Fig. 8 records an entire two-dimensional slice of the photoelectron distribution $E(\mathbf{k}_f)$ in front of the sample in a single measurement. For a Ag(100) surface, which is illuminated by photons with energy $h\nu = 35$ eV, such a slice is displayed as a colour-coded map in Fig. 11 for photoelectrons emitted around an angle $\theta = 20°$ with respect to the surface normal. The energy scale is renormalized to the Fermi energy E_F, and the photoelectron intensity is represented as a function of binding energy E_B and emission angle θ.

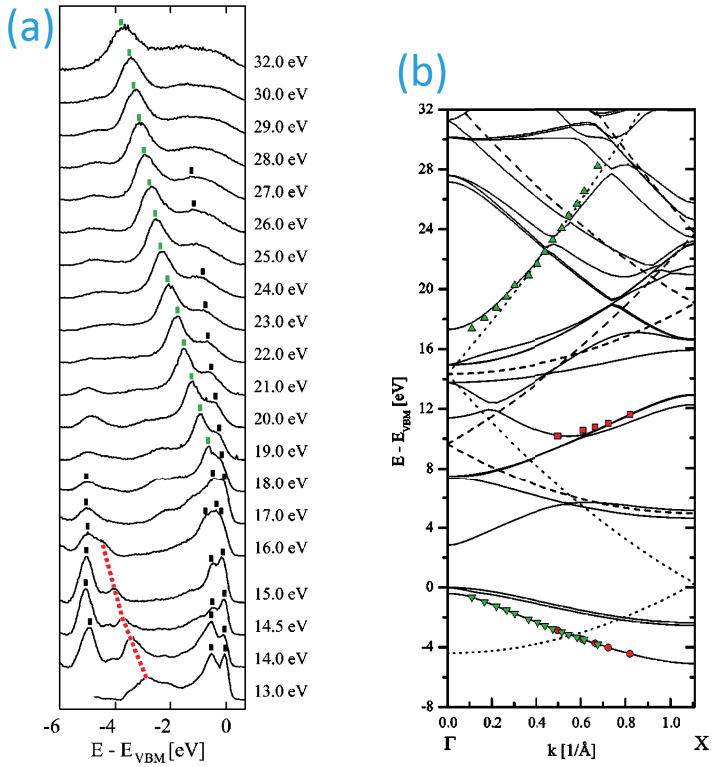


Fig. 12: *(a) Set of normal emission spectra of ZnSe(001) thin film measured at photon energies between hν = 13 eV and 32 eV. (b) Solid lines show theoretical occupied and unoccupied band structure of ZnSe along ΓX line in the Brillouin zone. Dashed lines show free electron bands. Red and green symbols show transitions related to the certain peak in the experimental spectra under the assumption of the correctness of the initial valence band dispersion.*

For a more detailed analysis of the spectral features one may take cuts through the $I(E_B, \theta)$ distribution, resulting in different types of spectra. A cut at fixed binding energy yields $I(\theta)$, which is sometimes called a *momentum distribution curve* (MDC). A cut at fixed angle yields $I(E_B)$, which is called an *energy distribution curve* (EDC) and corresponds to a "classical" photoemission spectrum.

Although the distribution in Fig. 11 seems to resemble a band structure, it is important to note that the data are not a simple picture of the bands, because the energy and angular position of the intensity maxima is determined by the transition matrix elements and thus by the initial state and final state bands. Nevertheless, we can already clearly discern different types of spectral features with large and smaller dispersion. In fact, a comparison to bulk band structure calculations of silver along the [100] (Δ) direction reveals that the spectral structure that starts at the Fermi level and bends downwards to the right originates from a strongly dispersing band of symmetry Δ_1, which has a strong free-electron, *sp*-like character. The more localized 4*d*-like states in silver give rise to the strong almost horizontal lines at binding energies below 4eV.

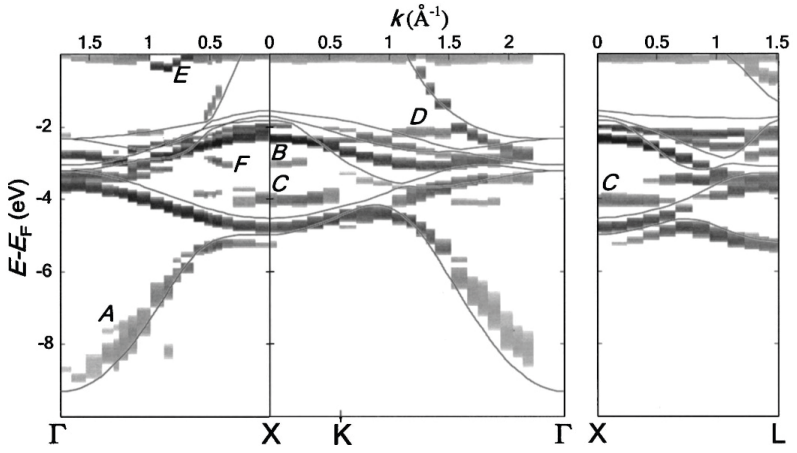


Fig. 13: *Band mapping results for the bulk electronic states in a Cu single crystal along the [100]* $(\Gamma - \Delta - X)$*, ([110]* $(\Gamma - \Sigma - K)$*, and [111]* $(\Gamma - \Lambda - L)$ *directions. The bands are shifted from the DFT theoretical* $E(\boldsymbol{k})$*, shown by thin lines, due to excited-state self-energy effects. The constant line at* E_F *is due to the Fermi cutoff, and the peaks A–F are spurious structures due to multiple upper band composition, 1DOS maxima, and surface states. From [36].*

For three-dimensional crystals significant band dispersion is present also in the direction perpendicular to the surface $k_\perp$. In order to investigate such effects it is the easiest to measure the set of normal emission spectra at the range of photon energies because the Eq. 17 simplifies to

$$k_\perp = \sqrt{\frac{2m}{\hbar^2}(E_{kin} + V_0)}. \tag{22}$$

Such set for the case of ZnSe(001) surface is presented in Fig. 12(a). Detailed analysis of these spectra is presented in [34] and in the following we will focus on the dominant feature in these spectra, indicated with the green marker, and another feature indicated with the red dashed line, which are interpreted in Fig. 12(b) (see also Fig. 7(b)). The feature marked in green closely follows the final band dispersion which is very close to the dispersion of the free-electron parabola. This means that in case of this feature Eq. 22 is a good approximation for finding the value of $k_\perp$. However, no matching free-electron final band can be found for the case of features indicated by red dashed line in Fig. 12(a). These features are not related to Auger transitions (their kinetic energy is not constant) and they are not related to surface states, because their initial (binding) energy is not constant. Therefore either they do not reflect the calculated initial band structure, or they are related to the transitions to non-free electron final bands. Figure 12(b) shows that there exist non-free-electron final bands which allow to interpret the dispersion of these features as originating from the same initial band as the features marked in green.

More sophisticated and accurate band mapping can be performed by employing experimental schemes for the determination of the final band dispersions from electron scattering measurements [35], and combining them with photoemission spectra. Fig. 13 displays such a result

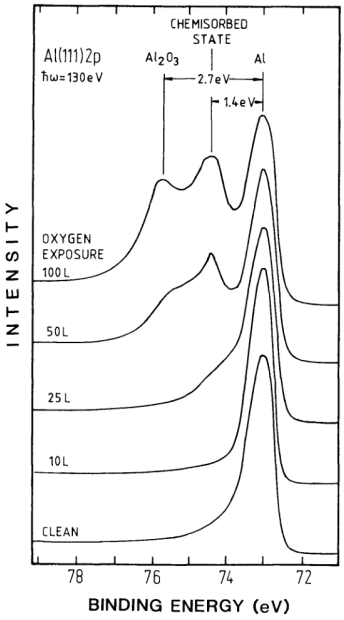


Fig. 14: *Core level photoemission from Al(111). A surface reaction with oxygen leads to characteristic chemical shifts of the core level binding energies with respect to the clean surface. The amount of oxygen that the surface is exposed to is measured in units of Langmuir L* ($1L = 10^{-6} mbar \cdot s$)*. From [1].*

for copper, with the data points being obtained from photoemission experiments and the lines representing a band structure calculation [37, 36]. The strongly dispersing bands starting at the Γ-point correspond to the free-electron like *sp*-type states. All other bands exhibit a weaker dispersion and have a strong *d*-type character, meaning that the electrons are more localized.
The band structure in Fig. 13 has been calculated within a relativistic scheme, i.e. it also contains the effects of spin-orbit interaction. Although copper is a material with low atomic number, spin-orbit coupling has been found to play an important role in the band symmetries, in particular, close to hybridization points. We also observe spectral signatures, which do not fit into the calculated bulk band structure ($A - F$) . A further analysis reveals that feature E is related to a surface state. The features B, C, and D are caused by transitions into regions with a strong one-dimensional density of states, where $k_\perp$ is not conserved. A hybridization of multiple upper bands leads to the appearance of feature A. Feature F is likely caused by surface state or surface resonance split off from the d bands [36]. We also note that the experimental data exhibit a systematic shift with respect to the calculated band. This is due to self-energy effects in the excited state.

4.1.2 Core level photoemission

The photoemission from the localized core levels gives rise to rather sharp spectral features at well-defined and characteristic binding energy values (see Fig. 3). These values are tabulated for the elements, for example, in the *X-Ray Data Booklet* [38], and range from several 10 eV for the shallow core levels up to 10 keV for the 1*s*-levels of heavy elements. Core level photoemission is often used to identify and quantify chemical species and is therefore also termed ESCA (Electron Spectroscopy for Chemical Analysis). For a sample consisting of one chemical element only, the binding energy of the spectral feature is sufficient to unambiguously

identify the element[3].

An example is given in Fig. 14, which compiles different spectra recorded for the Al 2*p* core level. The bottom most spectrum has been obtained from a clean Al(111) surface and shows the 2*p* core level peak located at a binding energy of $E_B = 73$ eV. A further inspection of the spectrum reveals that the spectral line has a finite width and an asymmetric shape. The line width is mainly determined by the lifetime of the core hole created in the excitation process and by the energy resolution of the electron spectrometer. The asymmetric line shape arises mainly due to the excitation of electron-hole pairs in the vicinity of the Fermi level. This corresponds to inelastic electron scattering of the photoelectron in the solid, effectively shifting some spectral weight to the low binding energy side of the peak. This asymmetric line shape may be modeled, for example with the Doniach-Sunjic approach [39].

The binding energy of a given core level may change, as soon as we alter the chemical environment, for example, by a chemical reaction. Although the chemical bonds formed with an atom as a consequence of the reaction involve mainly the valence electrons, they may cause a charge transfer from or to that atom. This process modifies the electrostatic screening in the atom, ultimately resulting in a slight shift of the core level binding energy. These so-called *chemical* shifts form the basis of more elaborate ESCA approaches in determining the chemical composition of complex alloys and compounds. An illustration for chemically induced core level binding energy shifts is given by the remaining spectra in Fig. 14. These spectra are obtained by exposing the Al(111) surface to different amounts of oxygen in successive steps. After an oxygen exposure of 25 L we start to see a weak spectral feature on the high binding energy side of the 2*p* level. After 50 L this feature has grown into a well-defined sharp peak $E_B = 74.4$ eV, which can be attributed to photoemission from Al surface atoms onto which oxygen has chemisorbed.

[3]Usually one measures several core level lines at different binding energies in order to increase the accuracy of the element analysis.

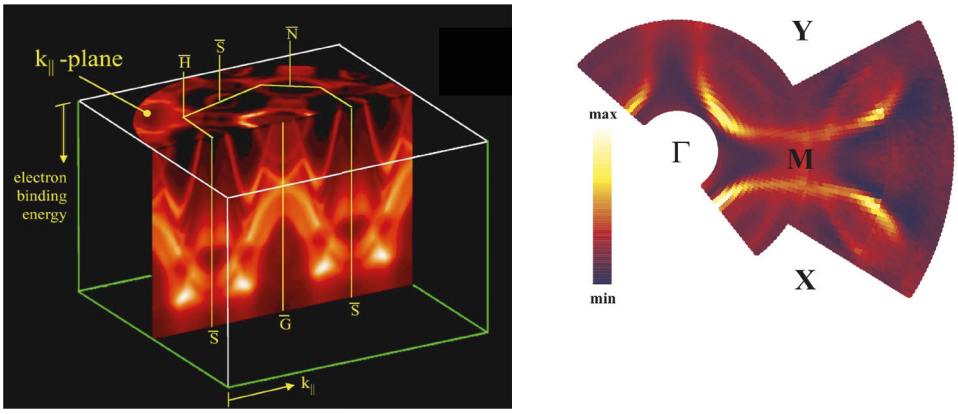


Fig. 15: *(Left) Principle of Fermi surface mapping illustrated for photoemission from W(110). The plot compiles photoemission intensity distributions $I(E, \boldsymbol{k}_\parallel)$ for different emission angles, which can be stacked in a three-dimensional scheme. A cut through the stack at $E = E_F$ yields a two-dimensional map of the Fermi surface in the plane defined by $\boldsymbol{k}_\parallel$. From [40]. (Right) Fermi surface for Pb-doped Bi2212, i.e. $Bi_2Sr_2CaCu_2O_{8-\delta}$. From [41].*

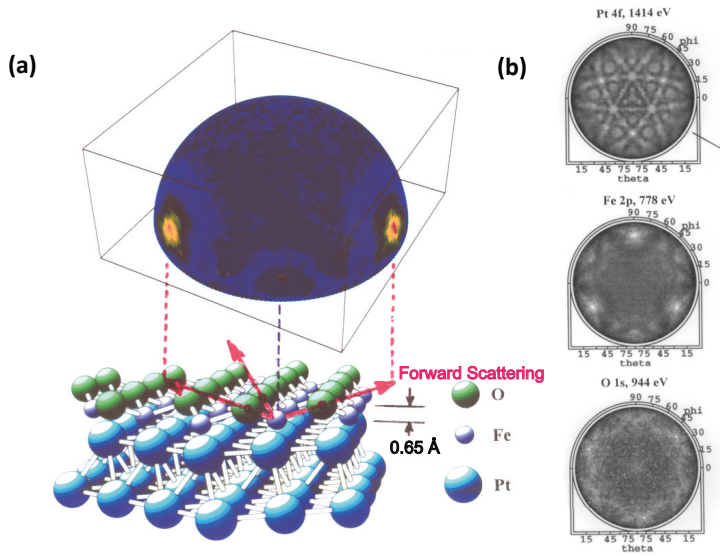


Fig. 16: *X-ray photoelectron diffraction at 1486.7 eV excitation from a monolayer of FeO grown on Pt(111). (a) A full-hemisphere pattern for Fe* $2p$ *emission is shown, above the atomic geometry finally determined for this overlayer. (b) Diffraction patterns simultaneously accumulated for emission from Pt* $4f$ *(kinetic energy 1414 eV), Fe* $2p$ *(778 eV), and O* $1s$ *(944 eV). From Ref. [42].*

In addition, a third peak starts to form at still higher binding energies. After dosing 100 L onto the Al(111) surface, this third signature at $E_B = 75.7$ eV has evolved into a clear peak, which can be attributed to photoemission from Al atoms bonded in an Al_2O_3 environment. We therefore see that the oxidation from metallic aluminium to alumina is accompanied by a chemical shift of the Al $2p$ core level by about $\Delta E_B = 2.7$ eV.

4.1.3 Fermi surface mapping

A particular aspect in modern photoemission spectroscopy is the so-called *Fermi surface mapping*. In order to see how this approach works, it is useful to recall that the data provided by the 2D display analyzers represent an intensity distribution $I(E, \mathbf{k}_\parallel)$. The electron wavevector $\mathbf{k}_\parallel$ parallel to the surface is defined by the experimental geometry. By varying the emission angles θ and ϕ (cf. Fig. 2) one obtains a set of slices through reciprocal space for different vectors $\mathbf{k}_\parallel = (k_x, k_y)$ in the surface plane. This data set can be condensed into a three-dimensional representation $E(k_x, k_y)$. An example for angle-resolved photoemission from a W(110) surface is shown in Fig. 15. The picture combines a vertical cut $I(E, k_x, k_y = 0)$ through the surface Brillouin zone (SBZ) with a horizontal cut $I(E = E_F, k_x, k_y)$. The intensity distribution $I(E, k_x, k_y = 0)$ reveals a clear dispersion of band segments along the high symmetry directions $\bar{\mathrm{S}} - \bar{\Gamma} - \bar{\mathrm{S}}$ in the SBZ, which can be compared to appropriate band structure calculations.

The horizontal cut $I(E = E_F, k_x, k_y)$ depicts a two-dimensional map of the electronic states at the Fermi energy and can thus be *related* to the Fermi surface. In the interpretation we have to keep in mind that the map in Fig. 15 contains matrix element and photoelectron diffraction effects. These have to be taken into account when comparing the data to theoretical predictions.

The details of the Fermi surface are crucial in determining the physical properties of materials, for example, the magnetic anisotropy in magnetic systems or the origin of superconductivity. In fact, the onset of superconductivity is accompanied by the formation of a small gap around the Fermi level. It is for this reason that Fermi surface mapping has become a standard tool in the investigation of high-T_C superconductors (HTSC). The example in Fig. 15 depicts the Fermi surface of Pb-doped $Bi_2Sr_2CaCu_2O_{8-\delta}$ (short form Bi2212) [41]. The data have been recorded in the normal state (T=120 K) with a He discharge source ($h\nu = 21.2$ eV). The main Fermi surface is hole-like and has the form of tubes (rings) centered around the X and Y high symmetry points. In addition, weaker intensity features are observed specifically around the M point. This so-called shadow Fermi surface is attributed to a spin-related origin [41]. A detailed understanding of the Fermi surface and their change with temperature are mandatory to understand the microscopic mechanisms leading to HTSC.

4.1.4 Photoelectron Diffraction

As one example of a photoelectron diffraction pattern, we show in Figure 16(a) the full hemisphere intensity distribution for Fe $2p$ emission at 778 eV ($\lambda_e = 0.44$ Å) from a monolayer of FeO grown on a Pt(111) surface [42]. At this energy, the forward-peaked nature of XPD is observed to create strong peaks in intensity along the Fe-O bond directions. The angle of these peaks can furthermore be used to estimate the distance between the Fe and O atoms in the overlayer, and it is found to be only about half that for similar bilayer planes in bulk FeO, as illustrated in the bottom of Figure 16(a). Figure 16(b) also illustrates the element-specific structural information available from XPD. The Pt $4f$ XPD pattern from the same sample is rich in structure due to the fact that emission arises from multiple depths into the crystal, with forward scattering producing peaks and other diffraction features along low-index directions. The Fe $2p$ pattern is here just a projection onto 2D of the 3D image in Figure 16(a). The O $1s$ pattern shows only very weak structure, as the O atoms are on top of the overlayer, with no forward scatterers above them, and only weaker back scattering contributing to the diffraction pattern. Comparing the Fe and O patterns thus immediately permits concluding that Fe is below O in the overlayer, rather than vice versa. Other examples of photoelectron diffraction in the study of clean surfaces, adsorbates, and nanostructure growth can be found elsewhere [21, 43].

4.2 Spin Effects in Photoemission

The electron spin can give rise to very peculiar phenomena in photoemission experiments. This is due to the fact that the electronic states are subject to two spin-dependent interactions: (i) spin-orbit coupling, and *(ii)* exchange interaction. Whereas spin-orbit coupling is mainly an atomic property, exchange-interaction is at the heart of the many electron system and is responsible for magnetic phenomena.

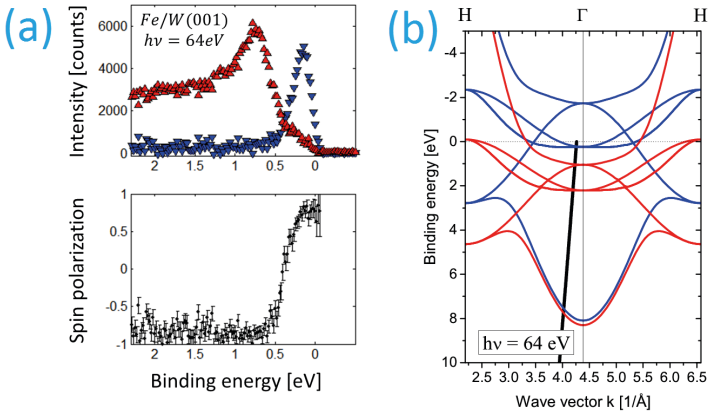


Fig. 17: *(a) Spin polarized spectra of Fe/W(001) taken at normal emission at hν = 64 eV and 100K using the Mott detector at the NSLS beamline U5 [44]. Sherman (asymmetry) function of $S = 0.17$ was used to calculate the plotted spectra, and the values on the ordinate scale of the upper panel results from the Eq. 21, however, they are close to the actual experimental accumulated count rate collected by the opposite channeltrons. (b) Bulk band structure of iron along ΓH line in the Brillouin zone, red/blue color indicates majority/minority exchange split bands. Black line shows the interpretation of the normal emission spectrum from Fe(001) at $h\nu = 64$ eV according to the Eq. 17.*

4.2.1 Ferromagnetic systems

A ferromagnet is characterized by a finite magnetization **M**, i.e. a spontaneous long-range magnetic order below a critical temperature T_C. The magnetization is related to a lifting of the spin-degeneracy of the valence electronic states. As a consequence, the spin-up and spin-down bands are separated in binding energy by the exchange splitting $\Delta E_{exc}(\mathbf{k}, E)$. A spin-resolved photoemission experiment will therefore be able to directly distinguish between the spin-up and spin-down states, as the spin is preserved during the optical transition.

Initially, ferromagnetic materials were often prepared as single crystals in the *picture frame* geometry [45], however, in such configurations stray magnetic fields of several mOe range are difficult to avoid. Such fields can influence the angular distribution of photoelectrons, in particular at lower kinetic energies, and therefore compromise the angular resolved measurements. Moreover, they can also affect the orientation of the spin polarization vector while the photoelectrons traverse the stray field region. To remove the influence of the stray fields nowadays ferromagnetic samples are typically prepared as epitaxial thin films, which readily deliver a single domain configuration when magnetized along one of the magnetic easy axes. The advantage of such a configuration is the possibility to perform a set of consecutive measurements with the film remanently magnetized in opposite directions, which enables one to effectively remove any spurious "apparatus" asymmetries introduced by the experimental setup.

When ferromagnetic epitaxial films with sufficient quality are prepared, such that the angular dependence of electrons reflects the reciprocal space with a minimal secondary electron background due to impurities and surface roughness, then clear features with a spin polarization close to 100% are observed in spin-ARPES spectra. Exchange interaction introduces a shift

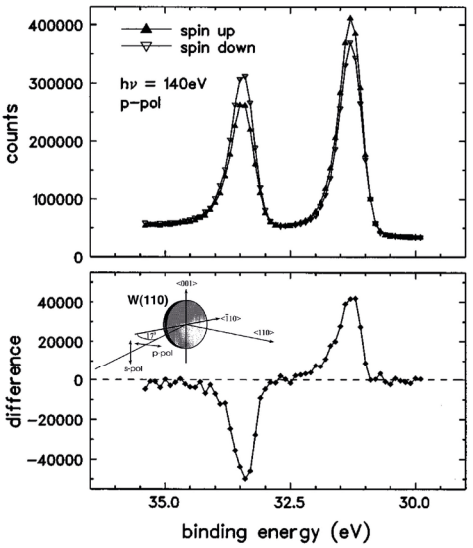


Fig. 18: *Spin-resolved W 4f energy distribution curves (EDC's) measured with p-polarized light of hν=140 eV photon energy. The spin polarization vector is oriented perpendicular to the plane spanned by the direction of light incidence and the surface normal. The integral of the difference (lower panel) over the binding energy vanishes within 1% relative to the integral of the absolute value of the difference. From [50].*

between majority and minority bands over the entire Brillouin zone, therefore spin polarization from ferromagnets is observed also in the angle integrated photoemission spectra. In case of high resolution studies another condition to obtain nearly 100% spin polarization, which is often neglected, is that the probed states originate from the Brillouin zone region, where no band splittings, resulting from lifted degeneracy of the crossing bands due to SOC (arising from the reduced symmetry due to the defined magnetization direction), take place.
Example spectrum which originates from exchange-split bands is presented in Fig. 17, where both the minority and majority features of Fe near the Γ are observed, showing a nearly full spin polarization. The two clear features result from d_{e_g} majority and $d_{t_{2g}}$ minority initial bulk bands of Fe, and their splitting of ~ 0.7 eV agrees well with ~ 0.8 eV predicted by *ab initio* calculations [46, 47].

4.2.2 Optical spin orientation

Even in the absence of magnetic interactions, it is possible to observe spin-polarized photoelectrons and relate them to the symmetry of the electronic states. This phenomenon is called "optical spin orientation" and the microscopic mechanism is provided by spin-orbit coupling, as we have already discussed in Sect. 2.1.1. The effects are large, if the spin-orbit coupling in the occupied states is strong.
As an example, Fig. 18 shows spin-resolved photoemission data for the W $4f$ shallow core levels obtained with linearly polarized light. The geometry was chosen such that the light impinges on the W(110) surface at a glancing angle of 17°. Symmetry arguments require that the spin-polarization vector is oriented perpendicular to the plane spanned by the direction of incidence an the surface normal [48]. These states show a clear spin-orbit splitting of about $\Delta E_{so} \approx 2.5$ eV between the $4f_{7/2}$ and $4f_{5/2}$. We see that the partial intensity spectra of spin-up (▲) and spin-down (∇) differ significantly at the peak positions, resulting in a positive spin polarization

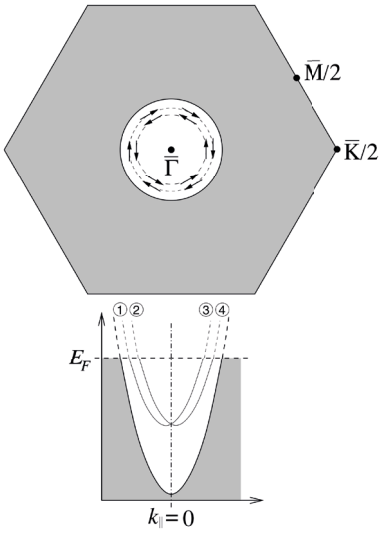


Fig. 19: *Upper panel, section of the surface Brillouin zone of the unreconstructed Au(111) surface. The $\overline{\Gamma K}$ distance is $\pi\sqrt{32}/3a = 1.45\mathring{A}^{-1}$. Lower panel, schematic view of the split surface state dispersion in a cut through $\overline{\Gamma}$. From [53].*

at the $4f_{7/2}$ emission line, whereas the $4f_{5/2}$ level exhibits a negative spin polarization (bottom panel). This spin polarization reversal between the spin-orbit split levels is an instrinsic feature of the optical spin orientation process, because the total spin polarization integrated over all spin-orbit split levels is required to vanish for symmetry reasons – at least in nonmagnetic materials. Note that for a given experimental geometry the sign of the spin polarization is unambiguously connected to the symmetry of the electron states involved in the optical transition. This assertion also holds for band states in a solid and allows a detailed analysis of spin-orbit effects in the band structure on the basis of spin-polarized photoemission experiments [49].

4.2.3 Rashba States

The spin-polarization effect described in the previous section is due to the intraatomic spin-orbit interaction included in the Hamiltonian (Eq. 2). This term has a specific structure. In the field of spin-dependent transport there is a strong desire to control the electron spin in semiconductors by electric fields. In order to describe this situation in a planar configuration, one often uses the Rasbha-Bychkov Hamiltonian [51]. Interestingly, it has a very similar mathematical form

$$H_{RB} = \alpha \boldsymbol{\sigma} \cdot (\mathbf{k} \times \mathbf{E}) \tag{23}$$

with the Rashba constant α, effective electric field $\mathbf{E}$, and Pauli matrices $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$. One expects maximal effects of the Rashba Hamiltonian when the electric field, the electron momentum and the electron spin are mutually orthogonal.
In two-dimensional (2D) systems with broken inversion symmetry, this spin–orbit interaction causes spin separation of the moving electrons – which is why it is interesting for spintronics. However, the inversion symmetry of the potential is also naturally broken at any crystal surface or interface. As a consequence, electronic states localized at a surface/interface should be spin-

split although this splitting can be quite small. In fact, the Rashba interaction at crystal surfaces becomes sizeable only when it couples to the large intra-atomic spin-orbit interaction. The gradient of the surface potential by itself is not sufficient to cause a directly observable splitting of the surface/interface electronic bands into spin subbands [52]. Therefore, this interaction plays an important role only if high-Z elements are involved at the surfaces or interfaces.

It is well-known that some noble metal surfaces exhibit pronounced surface states. This is also true for the unreconstructed Au(111) surface, which exhibits a Shockley-type surface state at the center of the surface Brillouin zone (SBZ), i.e. at the $\overline{\Gamma}$-point. This situation is depicted in Fig. 19. The surface state is characterized by a parabolic dispersion with $k_{||}$. Due to the Rashba interaction the surface state will spin-split, forming two concentric ring-shaped Fermi surfaces with opposite spin polarization in the SBZ.

Indeed this splitting can be clearly seen in a high-resolution photoemission experiment as shown Fig. 20. The gray-scale intensity map represents a slice through the Brillouin zone along the direction $\bar{\Gamma}\bar{\mathrm{K}}$. Without the Rashba interaction there would be only one parabolic trace centered around $k_{||} = 0$, corresponding to the dispersion of the surface state. The Rashba interaction introduces a symmetric splitting resulting in two parabolic traces with opposite spin polarization. The energy and momentum distribution curves show that the two traces are only degenerate at $k_{||} = 0$, but separated otherwise.

The (111) surfaces of silver and copper have very similar Shockley surface states, but much smaller predicted Rashba splitting which may be due to the smaller intraatomic spin-orbit coupling as mentioned above. Recently spin-orbit splitting in Cu(111) has been observed by ARPES experiments at very low photon energies (hν = 6 eV, laser-excited) [54]. These experiments set the current limit in momentum resolution of modern angle-resolved photoemission.

Spin polarization of Au(111) surface state can be observed by modern spin-polarized photoemission spectrometers, with recent progress achieved by the momentum microscope [55]. We refer to the lecture **A6** by C. M. Schneider for details of this study.

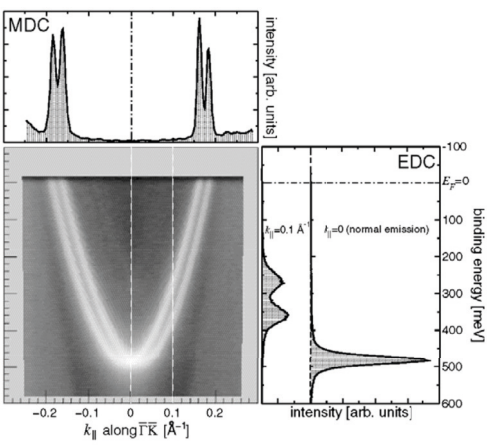


Fig. 20: *Photoemission intensity of the Shockley state on Au(111) as a function of energy and momentum $I(E_B, k_{||})$ (white means high intensity). The top panel shows a cut at constant energy $E = E_F$ (MDC); the right-hand panel gives the energy distribution curves (EDCs) at $k_{||} = 0$ and $k_{||} = 0.1\,\mathrm{\AA}^{-1}$. From [53].*

4.2.4 Topological Insulators

Recently a new class of materials called three-dimensional topological insulators (3D TIs) has been discovered and ARPES is one of the main experimental techniques used in their characterization [56]. Theoretical background on topological insulators is partly covered in this issue in the lecture **A2** by S. Blügel and G. Bihlmayer.

The most important 3D TIs are Bi_2Se_3, Bi_2Te_3, Sb_2Te_3 and their alloys. These compounds are narrow band gap semiconductors, which exhibit band character inversion at the Brilloin zone center Γ due to the spin-orbit coupling. This leads to the so-called *topologically nontrivial* phase in the bulk electronic structure of the material, as long as certain conditions for the parity of the bands at the time reversal invariant momenta points of the Brillouin zone are also matched [57, 58]. Such properties are described by the topological invariant Z_2, which is similar to the *genus* in topology. At the interface between the topologically nontrivial and topologically trivial material a robust interface state must exist. Since vacuum is topologically trivial, topological surface state must exist on every surface (which is an interface to vacuum) of a 3D TI, therefore 3D TIs are special materials which are insulating in the bulk, but have conducting surfaces.

Schematic comparison of topological and Rashba-type surface states is shown in Fig. 21 (b)-(c). In Rashba systems always even number of spin-polarized bands cut the Fermi level, while topological states consist of odd number of single-branched non-degenerate bands. Another related feature of topological states is spin-momentum locking, the orientation of the spin is perpendicular to the momentum **k**.

High resolution ARPES can directly image surface states of 3D TIs, and the example for the case of Bi_2Te_3 is shown in Fig. 21(e). It shows one difference compared to the model calculations: hexagonal warping of the Dirac cone. The warping is a result of the crystal structure symmetry and gets less pronounced closer to the Dirac point away from the bulk valence and conduction band edges. Not all 3D TIs exhibit clear warping, for example is it much less pronounced in Bi_2Se_3.

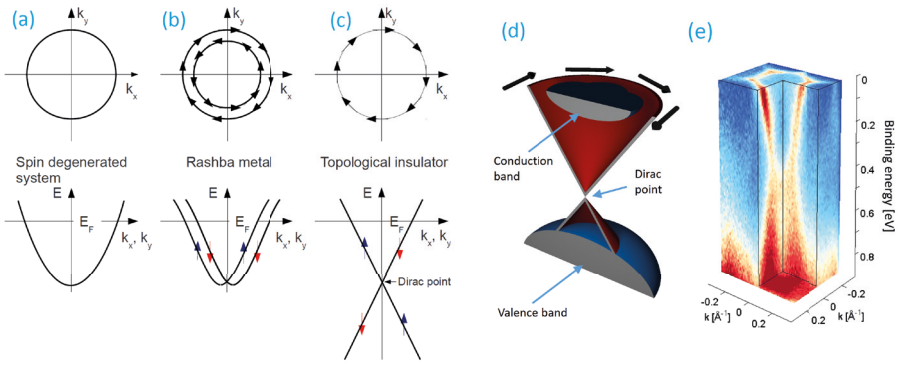


Fig. 21: *Schematic Fermi surface (top) and energy band dispersion (bottom) of a nonmagnetic material (a) without SO influence, (b) with SOC included Rashba spin splitting and (c) SOC induced single branched topological surface states. Panel (d) shows a schematic three-dimensional section through the Dirac cone showing the conduction band and the valence band, and panel (e) shows the actual ARPES measurement of the Dirac cone of Bi_2Te_3, which shows that the Fermi surface is not circular but undergoes a hexagonal warping.*

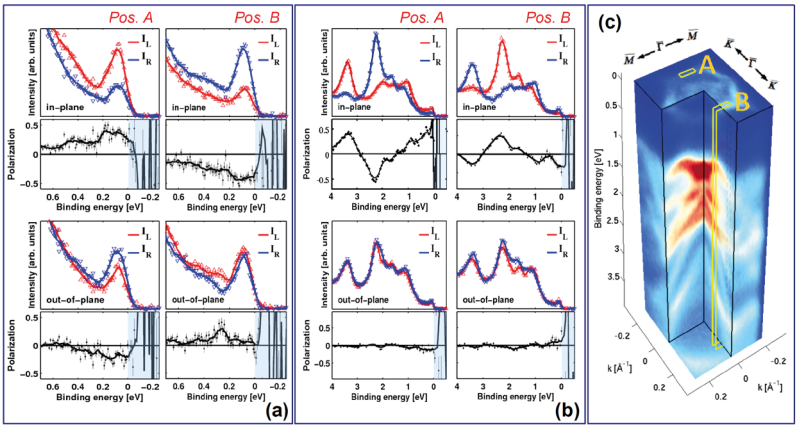


Fig. 22: *Spin-polarized data taken (a) near the Fermi level and (b) at higher binding energies on selected* ***k****-space locations along the* $\overline{\Gamma K}$ *direction on the 40 nm* Bi_2Te_3 *film [59]. Here, blue and red solid lines show smoothed* I_L *and* I_R *intensities, respectively. The top row indicates the in-plane spin-vector component intensities, whereas the out-of-plane intensities are plotted in the lower rows. The deduced spin-asymmetries* P_x *and* P_z *are plotted below the corresponding intensity plots with standard deviations given by vertical error bars, whereas the solid line represents smoothed data. (c) Experimental three-dimensional illustration of the band structure of a* Bi_2Te_3 *thin film over the full valence band region, indicating the* ***k****-space volumes A and B which are integrated in the spin-polarized experiment.*

Spin-momentum locking in topological surface states can be investigated by spin-polarized photoemission, and the results for Bi_2Te_3 are shown in Fig. 22. In Fig. 22(a)-(b) one can see that significant spin-polarization can only be observed for the in-plane spin component. Spin signal in the top panel of Fig. 22(a) reverses between measurement positions A and B which are marked in Fig. 22(c), which confirms spin-momentum locking. This reversal is also present in the wider range spectra in Fig. 22(b), which indicates the existence of other, Rashba-type spin polarized surface features in Bi_2Te_3. Small non-vanishing spin component in the out-of-plane spectra in the bottom panel of Fig. 22(a) is due to the hexagonal warping of the Dirac cone.

4.2.5 Magnetic Dichroism in Photoemission

What happens, if we have an experimental situation as described in sect. 4.2.2, but our sample is actually ferromagnetic? Let us take the example of a $2p$ core level. The ferromagnetic state is responsible for a spin-dependent energy splitting of the electronic states – not only in the valence states, but also in the core levels. These split according to their magnetic quantum number m_J, i.e. the $2p_{3/2}$ level splits into 4 sublevels ($m_{3/2}$, $m_{1/2}$, $m_{-1/2}$, $m_{-3/2}$), the $2p_{1/2}$ into two. The transition matrix elements depend on m_J *and* the orientation of the magnetization. As a consequence, the fine structure of the intensity spectrum depends on the magnetization direction. This phenomenon is called magnetic dichroism and is observed for both core levels and valence states [60].
This effect is shown in Fig. 23 for the $2p$ core level photoemission from Fe. Note that we have a very similar geometry as in experiment described in Sect. 4.2.2. The magnetization vector

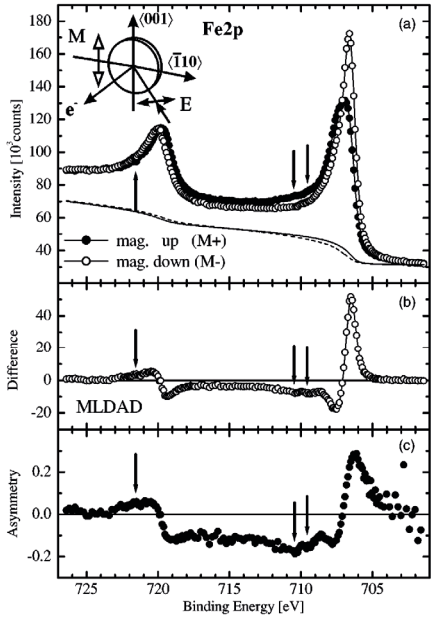


Fig. 23: *(a) Fe 2p photoemission spectra and Shirley background of 15 ML Fe / W(110) excited with p-polarized radiation ($h\nu$ =850 eV) for magnetization up and down (M+, M−). The inset shows the experimental geometry. (b) The intensity difference (MLDAD) of the curves from (a). (c) MLDAD asymmetry (without background). The arrows mark the position of correlation-induced satellites. From [61].*

is oriented perpendicular to the reaction plane. The upper panel compiles the photoemission spectra across the spin-orbit split $2p$ levels. We can see that the spectra differ significantly for opposite magnetization directions. The difference of the two spectra is plotted in the center panel and reveals characteristic bipolar signatures at the position of the core levels. We also note that the polarity of these features reverses between the $2p_{3/2}$ and the $2p_{1/2}$. This is consistent with the spin polarization change in the optical spin orientation experiment in Sect. 4.2.2. In fact, as a general rule, optical spin orientation phenomena in nonmagnetic materials are taking the form of magnetic dichroisms in ferromagnets.
The magnetic dichroism signal is often expressed as an intensity asymmetry A

$$A = \frac{I(M+) - I(M-)}{I(M+) + I(M-)}. \tag{24}$$

which reveals a similar spectral dependence compared to the difference. Additional weak spectral features are related to correlation effects (see Sect. 4.3). As the experiment has been performed with linearly polarized light, the effect is also termed magnetic linear dichroism in the photoelectron angular distribution (MLDAD). The latter points out that the size and sign of the magnetic dichroism depends strongly on the emission angle of the photoelectrons analyzed. A closer theoretical analysis shows, that the MLDAD is actually an interference effect between the two photoemission channels into s and d final states [62].

4.3 Electronic Correlations

It is well established nowadays that photoemission spectra of narrow-band materials, such as the elements of the d transition-metal series and their compounds, cannot be entirely explained

within a one-electron picture. This is due to the presence of local correlations between electrons in the partially filled d band. Experimental band mapping and its comparison with theoretical results can be a powerful tool to directly investigate correlation effects. It has to be realized, however, that the correlated electron picture is less transparent than the single particle model. The interactions due to the electronic correlations lead to a "dressing" of the single particle, i.e. when the particle moves in the solid it is always screened by these many-particle interactions. This system of particle and interaction cloud may be seen as a new quasiparticle. The respective many-electron calculations result in quasiparticle spectral functions rather than conventional band structures, which is a significant conceptual difference.

From all d transition metals, Ni has the narrowest bands and exhibits the strongest correlation effects. This can be seen in Fig. 24. Panel (a) reproduces a set of experimental angle-resolved photoemission spectra, which have been recorded for different emission angles from normal emission up to 70°, where several spectral features disperse with the emission angle. A comparison with a single particle particle calculation in the LDA approximation (panel b), however, predicts a much stronger dispersion of the bands than observed in the the experiment. In particular, strong spectral features should also be expected at binding energies larger than 0.5 eV. This is not observed in the experiment. Furthermore, the exchange splitting between bands of the same symmetry is calculated about twice as large, as observed in spin-resolved experiments ($\Delta E_{exc} \simeq 300\,\mathrm{meV}$ [64]).

The quasiparticle spectral functions calculated within a multiorbital Hubbard model for the experimental geometries are compiled in panel (c) of Fig. 24. The inclusion of correlation effects strongly modifies the spectra: all the structures are pushed up towards E_F by self-energy corrections reproducing much more closely the experimental results both in terms of energy position and dispersion. The spin dependence of the self-energy, arising from the different efficiencies of the scattering channels involving majority- and minority-spin electrons, strongly affects the spin polarization of the quasiparticle states. For this particular region in k space, four spin-up

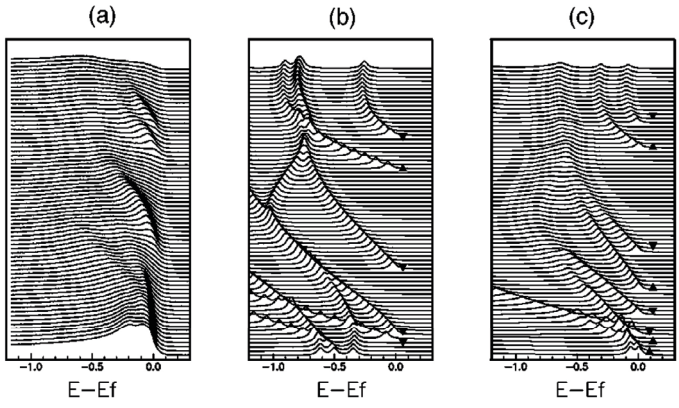


Fig. 24: *Comparison between (a) angle-resolved photoemission spectra from a Ni(110) surface at $h\nu = 21.2\,\mathrm{eV}$, (b) single particle local-density approximation (LDA), and (c) quasiparticle calculations results. The polar angle ranges from 0° (bottom) to 70° (top). The spin character is indicated by ▲ and ▼. From [63].*

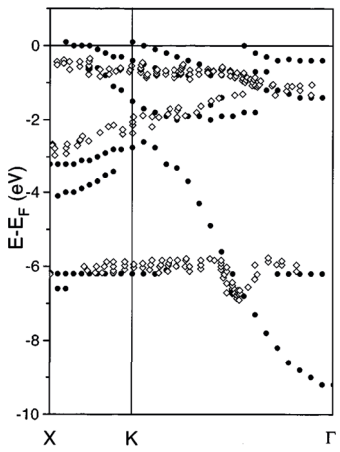


Fig. 25: *Comparison between the calculated dispersion of quasiparticle states (•) for majority-spin bands and angle- resolved spin-integrated photoemission results (◊) of Ref. [65]. From [66].*

and four spin-down bands are theoretically predicted in the energy region of interest. While in the single-particle picture one spin-up band and four spin-down bands cross the Fermi energy, all four spin-up bands come close to E_F after the inclusion of correlation effects. Moreover, the energy separation between the spin-up and spin-down bands between $\theta = 50°$ and $60°$ is reduced by self-energy corrections. All this is in excellent agreement with the experimental data.

In addition to a spin- and energy dependent renormalization of the quasiparticle states due to the self energy, the correlations also lead to the appearance of new spectral features, which are completely absent in the single particle band structures. The most prominent feature in Ni is the famous "6 eV satellite". This is depicted in Fig. 25, which shows the calculated dispersion of the majority spin quasiparticle states. These are compared to spin-integrated, angle-resolved photoemission results. In the region close to E_F we observe the spin-dependent energy renormalization already discussed above. In addition, we find a strong dispersing feature corresponding to the sp-type band. At about 6 eV below the Fermi level, however, there appears a new non-dispersing feature. This is the correlation-induced satellite, which indeed turns out to be of majority-spin character in spin-resolved photoemission experiments [67].

4.4 Kinkology

The on-site Coulomb interactions leading to the correlation phenomena discussed above are relatively strong and thus lead to large effects in the band structure. High-resolution photoemission nowadays provides the opportunity to study also the influence of much weaker interactions affecting the electronic system, for example, electron-phonon or electron-magnon interactions. As the analysis procedure is connected close to finding and identifying kinks and precisely measuring the spectral width in the dispersion of the quasiparticle states, this field is sometimes called "kinkology".

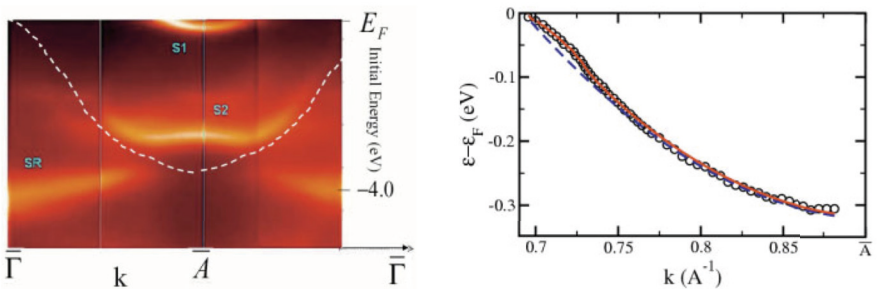


Fig. 26: *(Left panel) Energy vs. momentum photoemission display of the two surface state bands S1 and S2 on Be*$(10\bar{1}0)$*. The dashed line is the bulk band edge. Data taken at 30 K at 40 eV photon energy. (Right panel) Quasi-particle dispersion determined from momentum distribution curves (circles) obtained at 24 eV photon energy. Dashed blue line is the bare particle dispersion* $\varepsilon_0(k)$ *and the red line is the fit to the data from the extracted Eliashberg function. From [68].*

4.4.1 Electron-phonon interaction

The interaction of different quasiparticles, such as electrons and phonons, results in a crossing and hybridization of their respective dispersion relations. At the position in k-space where such crossings occur, the states involved are shifted in energy with respect to the noninteracting case. As phonons have very low energies of the order of 100 meV the respective modifications of the dispersion behavior of the electronic quasiparticle states due to the electron-phonon interaction will be confined to a narrow region below the Fermi level. Formally, the electron-phonon interaction can be considered as an additional contribution to the self energy Σ.
All characteristics of the electron-phonon coupling (EPC) are described by the *Eliashberg function* $E(\omega, \varepsilon, \mathbf{k}) = \alpha^2(\omega, \mathbf{k})F(\omega, \varepsilon, \mathbf{k})$, the total transition probability of a quasi-particle from/to the state $(\varepsilon, \mathbf{k})$ by coupling to phonon modes of frequency ω [69]. Information about the Eliashberg function can be obtained from the angle-resolved photoemission spectra, both through the EPC distortion of the quasi-particle bands near the Fermi energy and the temperature-dependent linewidth. If $\varepsilon_0(\mathbf{k})$ is the bare quasi-particle dispersion of a surface state without EPC, then the measured dispersion $\varepsilon(\mathbf{k})$ with electron-phonen coupling is given by

$$\varepsilon(\mathbf{k}) = \varepsilon_0(\mathbf{k}) + \mathrm{Re}\Sigma(\mathbf{k}, \varepsilon) \tag{25}$$

The screening of the electrons by the lattice is represented by the self-energy function $\Sigma(\mathbf{k}, \varepsilon)$. The imaginary part of the self-energy is related to the EPC contribution to the lifetime τ of the excited electronic states:

$$1/\tau = 2\mathrm{Im}\Sigma(\mathbf{k}, \varepsilon, T). \tag{26}$$

Based on these considerations the influence of the electron-phonon coupling has been investigated in Be$(10\bar{1}0)$ [68]. The photoemission data in Fig. 26 (left) show the dispersion of the surface states S1 and S2 as bright features. The experimental dispersion of the quasiparticle band $\varepsilon(\mathbf{k})$ (Fig. 26, right) is compared to the expected dispersion of the surface state without additional interactions $\varepsilon_0(\mathbf{k})$. This comparison reveals a weak, but distinct deviation of the ex-

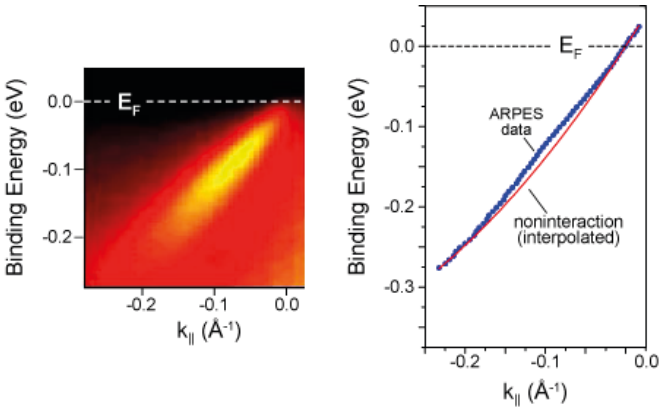


Fig. 27: *ARPES data from the iron (110) surface state. Left: Raw data, showing the intense quasiparticle region. Right: The electron band dispersion (E vs. $k_{||}$) extracted from the data reveals a weak "kink" in the region between 0.1 and 0.2 eV below E_F. From [70].*

perimental data from the parabolic dispersion close to the Fermi energy. This kink is the spectral signature of the electron-phonon coupling.

4.4.2 Electron-magnon interaction

In a magnet we have collective excitations of the spin system – magnons. These quasiparticles also have energies in the 100 meV range. We should therefore expect that electron-magnon interaction leads to the appearance of kinks in the band structure of ferromagnetic materials. This is demonstrated for the photoemission from the Fe(110) surface. The ARPES data (Fig. 27) show the spectral distribution of the surface state photoemission close to E_F at the center of the surface Brillouin zone. A careful analysis of surface state dispersion reveals a characteristic deviation from the parabolic behavior in the regime down to 200 meV below the Fermi level. This broader kink structure can be indeed related to the electron-magnon interaction [70]. From these data it is possible to extract the strength and extension of the electron magnon interaction.

4.5 High-Energy Photoemission (HAXPES)

So far we have discussed effects in valence band and core level photoelectron spectroscopy at excitation energies below 1000 eV. As we know from the inelastic mean free path curves under these conditions we will have $\lambda_{in} \simeq 1\mathrm{nm}$ at best, i.e. all of these experiments are surface sensitive (see Fig. 5). In recent years there is a strong effort to extend photoelectron spectroscopy also to higher excitation energies up to 10 keV in order to overcome this limitation. The approach is coined **H**Ard **X**-ray **P**hoto**E**lectron **S**pectroscopy (HAXPES) and poses several experimental challenges [71]. First, the electron spectrometers must be modified to be able to measure photoelectrons with high kinetic energy and good energy resolution ($\Delta E < 100\,\mathrm{meV}$). Second, the photoexcitation cross section for most core levels drops by 2-3 orders of magnitude, when going from 1 keV to 10 keV photon energy. As a consequence, the resulting photoelectron intensity will be small and difficult to measure. This can be only partially compensated on

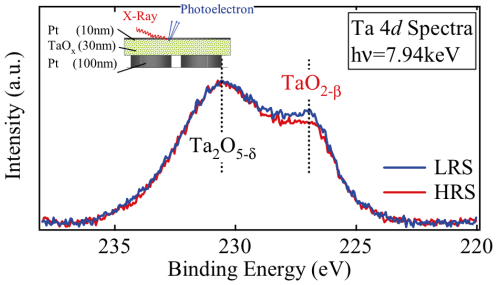


Fig. 28: *Hard X-ray photoemission spectra from Ta_2O_5 in the high (HRS) and low resistive state (LRS). From [73].*

the primary side, i.e. by increasing the photon flux. At present, HAXPES experiments are still demanding and very difficult to carry out with laboratory sources. With synchrotron radiation, however, HAXPES is quickly maturing into a powerful tool for materials characterization.

Core level analysis – The major advantage of HAXPES is its larger information depth which permits the access to buried layers and interfaces. The example shown in Fig. 31 is taken from the field of resistive oxides. Usually, oxides are wide band gap insulators. As is discussed in lecture **D6**, some of these materials may change their conductivity by several orders of magnitude, if a short current pulse above a certain threshold is applied to the material [72]. This current leads to the formation of conductive filaments or a local valency change in the oxide generating carriers for electrical transport. This is called the low resitive state (LRS). Interestingly, this process is reversible and the system may also be switched back into the high resistive state (HRS). This behavior considered as a future memory principle and explains the strong interest in resistive oxides.

Ta_2O_5 is one of the promising materials that has been investigated with respect to resistive memory applications. Fig. 31 shows the comparison of HAXPES spectra taken from the Ta $4d$ core states with about 8 keV photon energy. In order to switch the conductivity of the Ta_2O_5 film a bottom and top electrode usually made from Pt is needed, through which the switching current is passed through the insulator. This means, however, that the photoemission experiment has to probe the region below the Pt electrode, which requires a sufficiently high information depth. As can be seen, the experiment is indeed able to find a difference in the relative core level intensities underneath the 10 nm thick Pt-electrode, which can be related to a change of the oxidation state from Ta^{5+} to Ta^{4+} between the HRS and LRS state [73]. This demonstrates that HAXPES is able – at least in principle – to follow and map the valency changes taking place during the resistive switching process.

The second example relates to the field of spintronics. Magnetic tunneling barriers are considered as means to enable an efficient spin injection into semiconductors [74, 75]. One of the materials for spin-filter barriers investigated in this context is the ferromagnetic semiconductor EuO. In order to obtain a well-defined system for spin injection, a chemically and structurally sharp interface between EuO and the semiconductor – preferably silicon – must be established during the growth process. Of particular importance is the control of the oxygen partial pressure, as excess oxygen leads to a formation of interfacial silicon oxide.

The chemical quality of the EuO/Si interface can be addressed by HAXPES exploiting the kinetic energy dependence of the photoelectron inelastic mean free path (Fig. 29) [76]. The

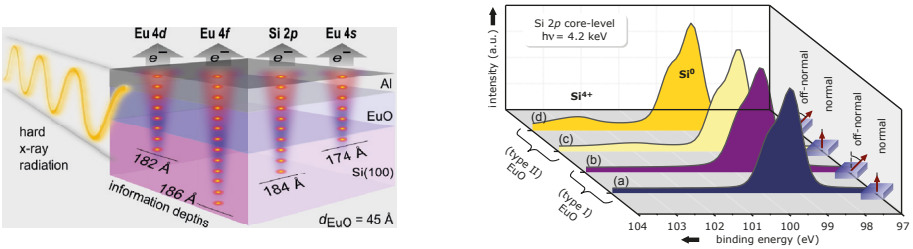


Fig. 29: *HAXPES on the EuO/Si interface. (Left) Schematic representation of the information depth for the different Eu and Si core levels. (Right) Si 2p core level photoemission spectra for (type I) stoichiometric EuO and (type II) O-rich EuO, recorded at 4.2 keV photon energy in normal (0°) and off-normal (60°) electron emission geometry. From [76].*

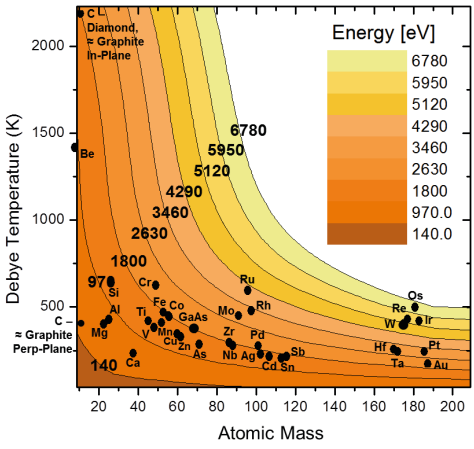


Fig. 30: *Contour plot showing Debye temperatures and photoelectron kinetic energies for the Debye-Waller factor W(T) = 0.5 at the sample temperature of 20K for various elements [78].*

thickness of the EuO film ($d_{EuO} = 45$ Å) has been chosen such that for a given photon energy (4.2 keV) the photoelectrons from the Si $2p$ levels reaching the spectrometer originate mainly from the interfacial region between EuO and Si. The interface sensitivity can be even increased by changing the take-off angle of the electrons from normal emission to off-normal emission. As can be seen in Fig. 29 the growth of *oxygen-rich* EuO (type II) leads to a significant photoemission satellite in the Si $2p$ spectrum which stems from a Si^{4+}state. The spectral weight of this contribution increases for the off-normal emission geometry. This is a clear indication that the silicon oxide contribution is located at the EuO/Si interface. The growth of *stoichiometric* EuO takes place at a lower oxygen partial pressure. The respective Si $2p$ spectra prove the absence of an oxide component, i.e. the interface between EuO and Si is chemically sharp. The results for the Eu $4d$, $4f$, and $4s$ core level photoemission corroborate the findings.

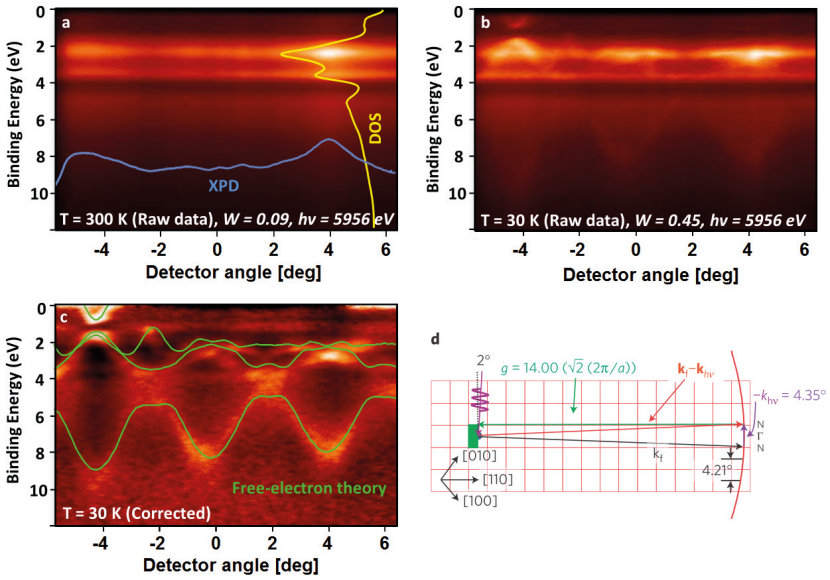


Fig. 31: *Temperature dependence of HARPES E(k) maps at hν = 6 keV [82]. (a) MEWDOS limit at room temperature with angular intensity modulations due to the X-ray photoelectron diffrations (XPD). (b) Clear signatures of band dispersions at T = 30K. (c) Comparison between background-corrected E(k) map and calculated band structure. (d) Extended BZ picture of the photoemission experiment at $h\nu = 6$ keV showing the related reciprocal lattice vector **G** and the photon momentum $\mathbf{k}_{h\nu}$.*

High energy angle-resolved photoemission (HARPES) – Angle-resolved photoelectron spectroscopy (ARPES) has been the method of choice to investigate the electronic band structure of crystalline surfaces and novel electronic materials. Traditional ARPES employs VUV photons ($20 - 150$ eV) which in the case of valence bands translates into similar kinetic energies of the emitted electrons. The inelastic mean free path (IMFP) of such electrons is $\lambda_{in} < 1$ nm, translating into a surface sensitivity, which is one of the key advantages of ARPES. On the downside, however, it obscures the access to the true bulk electron dispersion and to electronic states localized at buried interfaces.

One way of increasing λ_{in} and thus the probing depth of the ARPES experiment, is to employ low photon energies (i.e. below hν = 10 eV), however, this is material dependent, and the kinetic energy must be in any case larger than the work function W_F (for most surfaces W_F is around 4 to 5 eV). Therefore, the only safe way to increase λ_{in} is to perform HARPES (hard X-ray ARPES) experiments at photon energies in the multi-keV regime. IMFP curves for many elements are plotted in Fig. 5, and since λ_{in} should increase approximately as ${E_k}^{0.75}$, one can reach $\lambda_{in} \sim 30 - 60$ Å at 3-6 keV excitation energies. In addition a greater probing depth decreases the smearing in electron momentum perpendicular to the surface, which is proportional to $1/\lambda_{in}$ via the uncertainty principle.

Phonon effects set a fundamental limit of the momentum resolution of the HARPES experiment. At high temperatures signatures of the band dispersions fade out in $E(\mathbf{k})$ maps and the

valence band photocurrent reaches the matrix-element weighted density-of-states limit (MEWDOS, also often termed the XPS limit), typically also modulated by X-ray photoelectron diffraction effect [77]. Fraction of direct transitions can be estimated from the temperature-dependent Debye-Waller factor $W(T) \approx \exp\left(-G^2\langle u^2(T)\rangle\right)$, where **G** is the reciprocal lattice vector to allow direct transition, i.e. first Brillouin zone folding, and $\langle u^2(T)\rangle$ is the one-dimensional mean-squared vibrational displacement at temperature T. One can set $W = 0.5$ as a rather arbitrary limit at which realistic ARPES band mapping can be performed, and Fig. 30 shows $W(T)$ contours at 20 K for selected elements [78] where one can see, that band mapping at up to 2 keV is possible for most elements. It turns out that this is in most cases conservative, since clear signatures of dispersions can also be clearly observed for lower W values, in particular when suitable corrections for non-dispersive densities of states and photoelectron diffraction are applied to ARPES maps of suitable signal to noise ratio. Realistic simulations of the temperature-dependent HARPES spectra have been recently performed using the one-step photoemission formalism [79]. Another effect which cannot be neglected at the multi-keV energy range is the photoelectron energy loss due atomic recoil when a highly-energetic electron is emitted [80, 81].

The effect of temperature broadening is illustrated in Fig. 31(a-b) where $E(k)$ maps measured on W(110) surface at room temperature and at 30 K are presented [82]. These results were obtained at h$\nu = 6$ keV, which translates into $50 - 60$ Å probing depth, a true bulk sensitive band mapping. Debye-Waller factor for tungsten at 300 K is $W = 0.09$, therefore the room temperature spectrum in Fig. 31(a) shows the MEWDOS limit. At 30 K the Debye-Waller factor $W = 0.45$ and in Fig. 31(b) one can clearly see the signatures of tungsten bulk band dispersion.

At kinetic energies above ≈ 500 eV the final state of photoemission experiment can be approximated by free-electron parabola according to Eq. 15, which allows for convenient interpretation of the measured valence band dispersions. Such interpretation is presented in Fig. 31(c), where free-electron final-state ground state simulations based on a state-of-the-art density functional theory (DFT) with generalized gradient approximation (GGA) are found to be in a very good agreement with measured bands. Photon momentum $|\mathbf{k}_{h\nu}| = 2\pi\nu/c$, which is normally neglected in VUV ARPES, at multi-keV excitations reaches values comparable to the size of the Brillouin zone of a typical crystal. This is illustrated in Fig. 31(d), where photon momentum equals approximately the size of the entire Brillouin zone, which was taken into account in order to find the agreement shown in Fig. 31(c).

Recently HARPES has been used to characterize the details of the electronic structure of dilute mangetic semiconductor GaMnAs [83]. Further developments of HARPES technique will certainly involve improvements in experimental energy resolution and data acquisition times. Moreover, combining HARPES technique with the standing-wave ARPES [84] will allow probing the electronic state localize at specific depth below the surface.

4.6 Interfacial sensitivity

In Sect. 4.5 we have demonstrated two ways to vary the surface sensitivity in photoemission: changing the photon energy so as to move along curves of the type in Fig. 5 and varying the take-off angle, as indicated e.g. in Fig. 29. Both of these involve electron escape processes. One may also ask if there is a way to tailor the photon wave field so as to vary surface sensitivity. Creating an X-ray standing wave is one method for doing this, and it has been found possible to selectively look at buried layers and interfaces [85], as well as element-resolved densities of

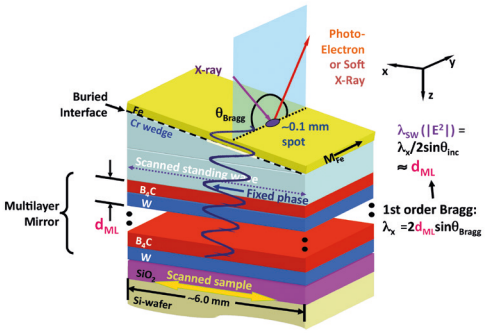


Fig. 32: *Schematic illustration of the simultaneous use of an X-ray standing wave plus a wedge-profile overlayer sample to selectively study buried interfaces and layers – the swedge method. In the example here, a strong standing wave (SW) is created by first-order Bragg reflection from a multilayer made of repeated B_4C/W bilayers, and a Cr wedge underneath an Fe overlayer permits scanning the SW through the Fe/Cr interface by scanning the sample along the x direction. From ref. [85].*

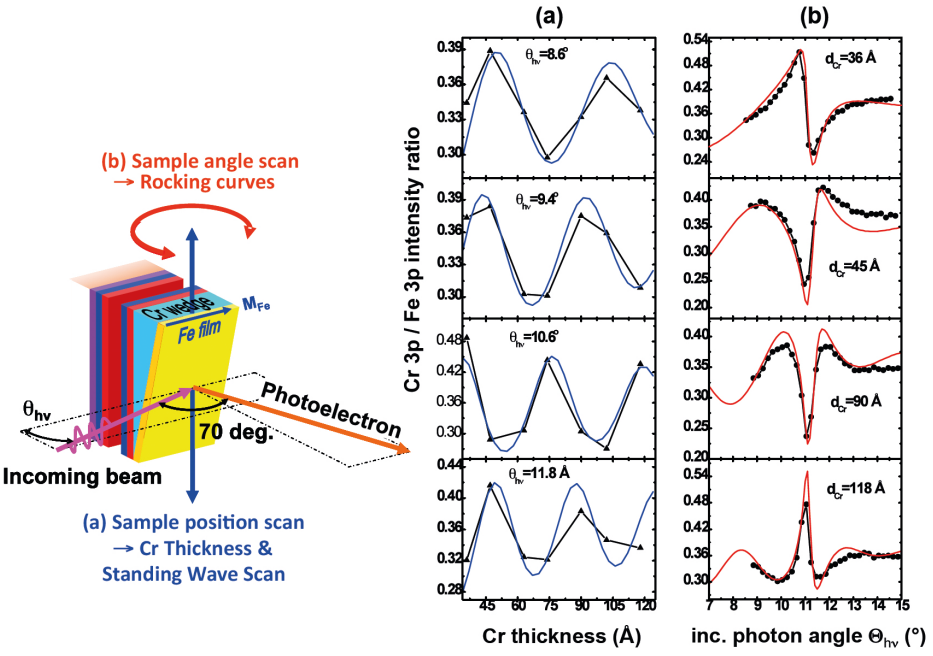


Fig. 33: *Experimental and calculated Cr 3p / Fe 3p ratios for two types of standing wave scan: (a) Scanning the sample along x at fixed incidence angle, as indicated in Fig. 32, and (b) scanning the sample polar angle with fixed x position (or Cr thickness). Also shown are best-fit theory curves. From Ref. [85].*

states [86], in this way.
In Fig. 32, we illustrate one approach for using soft X-ray (or in the future also hard X-ray) standing waves to carry out more precise depth-resolved photoemission from multilayer nanostructures [85]. This X-ray standing wave (XSW) approach combines a standing wave created by first-order Bragg reflection from a multilayer mirror of period d_{ML} with a sample in which one layer has a wedge profile, and can be termed the *swedge method*. If the standing wave is created by a typically well-focussed synchrotron radiation beam, then its dimensions will be much smaller than a typical sample, as indicated in the figure. Since the standing wave only exists in the region where the beam hits the sample surface, and its phase is locked tightly to the multilayer mirror, scanning the sample in the photon beam along the x direction effectively translates the standing wave through the sample. In the example shown, the standing wave would in particular scan through the Fe/Cr interface of interest, at some positions being more sensitive to the Fe side and at some more sensitive to the Cr side.
Some results obtained with this method for the Fe/Cr interface are summarized in Figs. 33 and 34. The analysis combined XPS intensity and MCDAD measurements (not shown here) from the $3p$ and $2p$ core levels of Fe and Cr, respectively. In Fig. 33(a) is shown the variation of the Cr $3p$ / Fe $3p$ ratio as the sample is scanned in the way suggested above, for several angles of incidence near the Bragg angle. Oscillations in this ratio clearly reflect the passage of the standing wave node and belly through the interface. In Fig. 33(b) we compile *rocking curves* in which the angle is varied around the Bragg angle for different positions x along the sample, or equivalently different Cr wedge thickness d_{Cr}. Also in this data there are sizeable changes in the intensity ratio.
Self-consistently analyzing these data with X-ray optical calculations of standing-wave photoemission and only two variable parameters (the depth of onset of change in the Fe composition and the width of a linear gradient as the interface changes from pure Fe to pure Cr) yields the excellent fits shown to both types of data, and the parameters given at the left side of Fig. 34(a). The MCDAD data for both Fe $2p$ and Cr $2p$ core level photoemission have also been measured as the sample is scanned in the beam. The relative signs of the MCDAD signal for the Fe $2p$ and Cr $2p$ levels are found to be opposite [85]. This immediately implies that a small amount of Cr is oppositely magnetized compared to Fe, which is induced by the ferromagnetic Fe layer, since Cr is normally antiferromagnetic. Similar data have been obtained at the $3p$ levels of Cr and Fe. Further analyzing this data set with two parameters for Fe $2p$ and $3p$ MCD and two parameters for Cr $2p$ and $3p$ MCD yields the atom-specific magnetization profiles shown at right hand side of Fig. 34(a).
Thus, in the above described experiment the *swedge method* has permitted non-destructively determining the concentration profile through an interface, as well as the atom-specific magnetization contributions through it. The *swedge* approach has also been used successfully to determine layer-specific densities of states that can be linked to changes in magnetoresistance as a function of nanolayer thicknesses [87]. Several other possible applications of it have also been suggested [85, 88, 89], including going to hard X-ray excitation, for which reflectivities and thus standing wave strengths can be much higher.

5 Conclusions

In this contribution, we could only touch upon selected aspects of photoelectron spectroscopy and photoemission processes. It should have become clear that this spectroscopy with its many

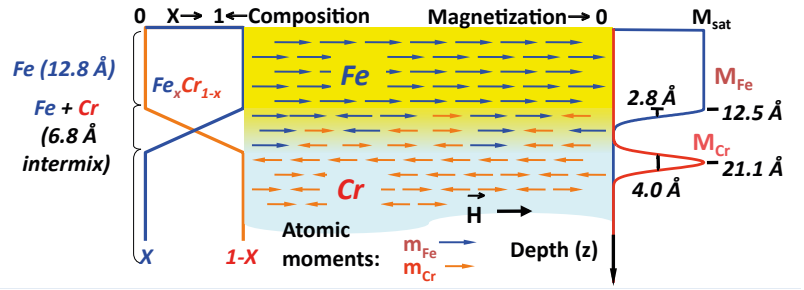


Fig. 34: *The concentration and atom-specific magnetization profiles through the Fe/Cr interface, as derived from the XPS and MCDAD experiments. From ref. [85].*

facets is a powerful tool for electronic and chemical characterization of materials. Very important information can already be extracted by means of qualitative interpretation schemes. The full potential, however, can be unleashed by quantitative descriptions within sophisticated photoemission calculations. The successful expansion of photoemission techniques to hard X-ray excitation relaxes the constraint of surface sensitivity. HAXPES offers access to genuine bulk electronic structures and buried interfaces.

Acknowledgement

This manuscript is in a large part based on the excellent contribution by C. M. Schneider from the 43rd IFF Spring School 2012.
The author is indebted to the Jülich spectroscopy and microspectroscopy groups at the storage ring facilities DELTA (Dortmund), BESSY (Berlin) and ELETTRA (Trieste). Sincere thanks are due to C. S. Fadley (Lawrence Berkeley National Laboratory), A. X. Gray (Temple University, Philadelphia), and E. Vescovo (NSLSII, Brookhaven National Laboratory) for ongoing collaborations and the permission to use material for this lecture.

References

[1] S. Hüfner, *Photoelectron Spectroscopy* 3rd ed. (Springer, Berlin, 2003).

[2] *Solid-State Photoemission and Related Methods*, eds. W. Schattke, and M. A. van Hove (Wiley-VCH, Weinheim, 2003).

[3] F. Reinert and S. Hüfner, *Photoemission spectroscopy-from early days to recent applications*, New J. Phys. **7**, 97 (2005).

[4] *Very High Resolution Photoelectron Spectroscopy*, ed. S. Hüfner (Springer, Berlin, 2007).

[5] F. de Groot and A. Kotani, *Core Level Spectroscopy of Solids* (CRC Press, Boca Raton, 2008).

[6] S. Suga and A. Sekiyama, *Photoelectron Spectroscopy, Bulk and Surface Electronic Structures*, Springer Series in Optical Sciences (Springer, Berlin, Heidelberg, 2014).

[7] C. M. Schneider, C. Wiemann, M. Patt, V. Feyer, L. Plucinski, I. P. Krug, M. Escher, N. Weber, M. Merkel, O. Renault, N. Barrett, *Expanding the view into complex material systems: From micro-ARPES to nanoscale HAXPES*, Journal of Electron Spectroscopy and Related Phenomena 185, 330 (2012).

[8] H. Hertz, *Über einen Einfluss des ultravioletten Lichtes auf die electrische Entladung*, Annal. Phys. **267**, 983 (1887).

[9] A. Einstein, *Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt*, Annal. Phys. **322**, 132 (1905).

[10] A. H. Compton, *A Quantum Theory of the Scattering of X-rays by Light Elements*, Phys. Rev. 21, 483 (1923).

[11] C. Nordling, E. Sokolowski, and K. Siegbahn, *Precision Method for Obtaining Absolute Values of Atomic Binding Energies*, Phys. Rev. **105**, 1676 (1957).

[12] *Handbook on Synchrotron Radiation* Vol. 1A, B, edited by D. Eastman and Y. Farge (North-Holland Publishing, Amsterdam, 1983).

[13] D. E. Starr, Z. Liu, M. Hävecker, A. Knop-Gericke and H. Bluhm, *Investigation of solid/vapor interfaces using ambient pressure X-ray photoelectron spectroscopy*, Chem. Soc. Rev. **42**, 5833 (2013).

[14] R. M. Martin, *Electronic Structure – Basic Theory and Practical Methods* (Cambridge University Press, Cambridge, 2004).

[15] K. Horn, in *Handbook of Surface Science*; Vol. 2, edited by K. Horn and M. Scheffler (Elsevier, Amsterdam, 2000), p. 383.

[16] *Optical Orientation*, edited by F. Meier and B. P. Zakharchenya (North-Holland, Amsterdam, 1984).

[17] H. Ebert, J. Minar, and V. Popescu, in: *Band Ferromagnetism*; edited by K. Baberschke, M. Donath, and W. Nolting (Springer-Verlag, Berlin, 2001), p. 371.

[18] R. N. Zare, *Angular Momentum* (Wiley, New York, 1988).

[19] D. T. Pierce and F. Meier, *Photoemission of spin-polarized electrons from GaAs*, Phys. Rev. B **13**, 5484 (1976).

[20] S. Tanuma, C. J. Powell, and D. R. Penn, *Calculations of electron inelastic mean free paths : VIII. Data for 15 elemental solids over the 50-2000 eV range*, Surf. Interface Anal. **37**, 1 (2005); S. Tanuma, C. J. Powell, and D. R. Penn, *Calculations of electron inelastic mean free paths. IX. Data for 41 elemental solids over the 50 eV to 30 keV range*, Surf. Interface Anal. **43**, 689 (2011).

[21] C. S. Fadley, in: *Synchrotron Radiation Research: Advances in Surface and Interface Science*, R. Z. Bachrach, Ed. (Plenum Press, New York, 1992).

[22] Multiple scattering program for calculating photoelectron diffraction available at: http://csic.sw.ehu.es/jga/software/edac/index.html, with the methodology behind it described in F.J. Garcia de Abajo, M.A. Van Hove, and C.S. Fadley, Phys. Rev. B **63**, 075404 (2001).

[23] R. Feder, in: *Polarized Electrons in Surface Physics*, ed. by R. Feder (World Scientific, Singapore, 1985).

[24] J. Braun, *The theory of angle-resolved ultraviolet photoemission and its applications to ordered materials*, Rep. Prog. Phys. 59 (1996).

[25] G. Borstel, *Theoretical Aspects of Photoemission*, Appl. Phys. A **38**, 193 (1985).

[26] A. Damascelli, Z. Hussain, and Z.-X. Shen, *Angle-resolved photoemission studies of the cuprate superconductors*, Rev. Mod. Phys. **75**, 473 (2003).

[27] A. Oelsner, O. Schmidt, M. Schicketanz, M.J. Klais, G. Schönhense, V. Mergel, O. Jagutzki, H. Schmidt-Böcking, *Microspectroscopy and imaging using a delay line detector in time-of-flight photoemission microscopy*, Rev. Sci. Instrum. **72**, 3968 (2001).

[28] L. Plucinski, A. Oelsner, F. Matthes, and C.M. Schneider, *A hemispherical photoelectron spectrometer with 2-dimensional delay-line detector and integrated spin-polarization analysis*, J. Elec. Spectroscopy 181, 215 (2010).

[29] N. Takahashi, F. Matsui, H. Matsuda, Y. Hamada, K. Nakanishi, H. Namba, and H. Daimon, *Improvement of display-type spherical mirror analyzer for real space mapping of electronic and atomic structures*, J. Electron Spectr. Rel. Phen. **163**, 45 (2008).

[30] J. Kessler, *Polarized Electrons*, 2nd ed. (Springer-Verlag, Berlin, 1985).

[31] J. Kirschner, *Polarized Electrons at Surfaces*, Springer Tracts in Modern Physics Vol. 106 (Springer-Verlag, Berlin, 1985).

[32] H. Siegbahn and K. Siegbahn, *ESCA applied to liquids*, J. Electron Spectrosc. Relat. Phenom. **2**, 319 (1973); H. Siegbahn, *Electron spectroscopy for chemical analysis of liquids and solutions*, J. Phys. Chem. **89**, 897 (1985).

[33] M. Müller, S. Nemsak, L. Plucinski, and C. M. Schneider, *Functional Materials for Information and Energy Technology: Insights by Photoelectron Spectroscopy*, J. Elec. Spectroscopy, in press, doi:10.1016/j.elspec.2015.08.003 (2015), and references therein.

[34] L. Plucinski, R. L. Johnson, A. Fleszar, W. Hanke, W. Weigand, C. Kumpf, C. Heske, E. Umbach, T. Schallenberg and L. W. Molenkamp, *Valence band electronic structure of ZnSe(001): Theory and Experiment*, Phys. Rev. B **70**, 125308 (2004).

[35] V. N. Strocov, *Low energy electron reflection: Possibility for E(k) points mapping above the vacuum level*, Solid State Commun. 78, 845 (1991).

[36] V. N. Strocov, R. Claessen, G. Nicolay, S. Hüfner, A. Kimura, A. Harasawa, S. Shin, A. Kakizaki, H. I. Starnberg, P. O. Nilsson, and P. Blaha, *Three-dimensional band mapping by angle-dependent very-low-energy electron diffraction and photoemission: Methodology and application to Cu*, Phys. Rev. B **63**, 205108 (2001).

[37] V. N. Strocov, R. Claessen, G. Nicolay, S. Hüfner, A. Kimura, A. Harasawa, S. Shin, A. Kakizaki, P. O. Nilsson, H. I. Starnberg, and P. Blaha, *Absolute Band Mapping by Combined Angle-Dependent Very-Low-Energy Electron Diffraction and Photoemission: Application to Cu*, Phys. Rev. Lett. **81**, 4943 (1998).

[38] Available as pdf-file at `http://xdb.lbl.gov/`

[39] S. Doniach and M. Sunjic, *Many-electron singularity in X-ray photoemission and X-ray line spectra from metals*, J. Phys. C **3**, 285 (1970).

[40] E. Rotenberg, Advanced Light Source Berkeley, (priv. communication).

[41] S. V. Borisenko, M. S. Golden, S. Legner, T. Pichler, C. Dĭr, M. Knupfer, J. Fink, G. Yang, S. Abell, and H. Berger, *Joys and Pitfalls of Fermi Surface Mapping in Bi2Sr2CaCu2O8+d Using Angle Resolved Photoemission*, Phys. Rev. Lett. **84**, 4453 (2000).

[42] Y. J. Kim, C. Westphal, R. X. Ynzunza, Z. Wang, H. C. Galloway, M. Salmeron, M. A. Van Hove, C. S. Fadley, *The growth of iron oxide films on Pt(111): a combined XPD, STM, and LEED study*, Surf. Sci. **416**, 68 (1998).

[43] J. Osterwalder, A. Tamai, W. Auwarter, M.P. Allan, and T. Greber, *Photoelectron Diffraction for a Look inside Nanostructures*, Chimia **60** (2006) A795, and earlier references therein.

[44] L. Plucinski, Yuan Zhao, C.M. Schneider, B. Sinkovic, and E.Vescovo, *Surface electronic structure of ferromagnetic Fe(001)*, Phys. Rev. B 80, 184430 (2009).

[45] E. Kisker, R. Clauberg, and W. Gudat, *Electron spectrometer for spin-polarized angle- and energy-resolved photoemission from ferromagnets*, Rev. Sci. Instrum. 53, 1137 (1982).

[46] L. Plucinski, Yuan Zhao, E. Vescovo, and B. Sinkovic, *MgO/Fe(001) interface: A study of the electronic structure*, Phys. Rev. B 75, 214411 (2007).

[47] F. Matthes, L.-N. Tong, and C.M. Schneider, *Spin-polarized photoemission spectroscopy of the MgO/Fe interface on GaAs(100)*, J. Appl. Phys. 95, 7240 (2004).

[48] E. Tamura and R. Feder, *Spin Polarization in Normal Photoemission by Linearly Polarized Light from Non-Magnetic (110) Surfaces*, Europhys. Lett. **16**, 695 (1991).

[49] C. M. Schneider and J. Kirschner, *Spin- and angle-resolved photoelectron spectroscopy from solid surfaces with circularly polarized light*, Crit. Rev. Solid State Mater. Sci. **20**, 179 (1995).

[50] H. B. Rose, A. Fanelsa, T. Kinoshita, Ch. Roth, F. U. Hillebrecht, and E. Kisker, *Spin-orbit-induced spin polarization in W 4f photoemission*, Phys. Rev. B **53**, 1630 (1996).

[51] Y. A. Bychkov and E. I. Rashba, *Oscillatory effects and the magnetic-susceptibility of carriers in inversion-layers*, J. Phys. C: Solid State Phys. **17**, 6039 (1984), Y. A. Bychkov and E. I. Rashba, *Properties of a 2D eelctron-gas with lifted spectral degeneracy*, Sov. Phys. JETP Lett **39**, 78 (1984).

[52] S. LaShell, B. A. McDougall and E. Jensen, *Spin Splitting of an Au(111) Surface State Band Observed with Angle Resolved Photoelectron Spectroscopy*, Phys. Rev. Lett. **77**, 3419 (1996).

[53] F. Reinert, *Spin-orbit interaction in the photoemission spectra of noble metal surface states*, J. Phys.: Condens. Matt. **15**, S693 (2003).

[54] A. Tamai, W. Meevasana, P. D. C. King, C. W. Nicholson, A. de la Torre, E. Rozbicki, and F. Baumberger, *Spin-orbit splitting of the Shockley surface state on Cu(111)*, Phys. Rev. B **87**, 075113 (2013).

[55] C. Tusche, A. Krasyuk, and J. Kirschner, *Spin resolved bandstructure imaging with a high resolution momentum microscope*, Ultramicroscopy **159**, 520 (2015).

[56] M. Z. Hasan and C. L. Kane, *Colloquium: Topological insulators*, Rev. Mod. Phys. **82**, 3045 (2010).

[57] Liang Fu, C. L. Kane, and E. J. Mele, *Topological Insulators in Three Dimensions*, Phys. Rev. Lett. **98**, 106803 (2007).

[58] H. Zhang, C.-X. Liu, X.-L. Qi, X. Dai, Z. Fang, and S.-C. Zhang, *Topological insulators in Bi2Se3, Bi2Te3 and Sb2Te3 with a single Dirac cone on the surface*, Nature Physics **5**, 438 (2009),

[59] A. Herdt, L. Plucinski, G. Bihlmayer, G. Mussler, S. Döring, J. Krumrain, and D. Grützmacher, S. Blügel, and C. M. Schneider, *On the nature of the spin polarization limit in the warped Dirac cone of the Bi2Te3*, Phys. Rev. B **87**, 035127 (2013).

[60] W. Kuch and C. M. Schneider, *Magnetic dichroism in valence band photoemission*, Rep. Prog. Phys. **64**, 205 (2001).

[61] C. Bethke, E. Kisker, N. B. Weber, and F. U. Hillebrecht, *Core-valence interactions in Cr and Fe 2p photoemission*, Phys. Rev. B **71**, 024413 (2005).

[62] D. Venus, *Magnetic circular dichroism in angular distributions of core-level photoelectrons*, Phys. Rev. B **48**, 6144 (1993).

[63] F. Manghi, V. Bellini, J. Osterwalder, T. J. Kreutz, P. Aebi, and C. Arcangeli, *Correlation effects in the low-energy region of nickel photoemission spectra*, Phys. Rev. B **59**, R10409 (1999).

[64] K. Ono, K. Shimada, Y. Saitoh, T. Sendohda, A. Kakizaki, T. Ishii, and K. Tanaka, *Spin- and k-dependent electronic structure of ferromagnetic nickel*, J. Electron Spectr. Rel. Phen. **78**, 325 (1996).

[65] Y. Sakisaka, T. Komeda, M. Onchi, H. Kato, S. Masuda, and K. Yagi, *Photoemission study of the valence-band satellite of Ni(110)*, Phys. Rev. B **36**, 6383 (1987).

[66] F. Manghi, V. Bellini, and C. Arcangeli, *On-site correlation in valence and core states of ferromagnetic nickel*, Phys. Rev. B **56**, 7149 (1997).

[67] R. Clauberg, W. Gudat, E. Kisker, E. Kuhlmann, and G. M. Rothberg, *Nature of the Resonant 6-eV Satellite in Ni: Photoelectron Spin-Polarization Analysis*, Phys. Rev. Lett. **47**, 1314 (1981).

[68] S.-J. Tang, J. Shi, B. Wu, P. T. Sprunger, W. L. Yang, V. Brouet, X. J. Zhou, Z. Hussain, Z.-X. Shen, Z. Zhang, and E. W. Plummer, *A spectroscopic view of electron-phonon coupling at metal surfaces*, phys. stat. sol. (b) **241**, 2345 (2004).

[69] G. Grimvall, *The Electron-Phonon Interaction in Metals*, Selected Topics in Solid State Physics, edited by E. Wohlfarth (North-Holland, New York, 1981).

[70] M. Escher, N. Weber, M. Merkel, C. Ziethen, P. Bernhard, G. Schönhense, S. Schmidt, F. Förster, F. Reinert, B. Krömker, and D. Funnemann, *NanoESCA: a novel energy filter for imaging X-ray photoemission spectroscopy*, J. Phys.: Condens. Matt. **17**, S1329 (2005).

[71] K. Kobayashi, *Hard X-ray photoemission spectroscopy*, Nucl. Instrum. Methods A **601**, 32 (2009).

[72] R. Waser and M. Aono, *Nanoionics-based resistive switching memories*, Nat. Mater. **6**, 833 (2007).

[73] H. Kumigashira (priv. communication).

[74] R. Meservey and P. M. Tedrow, *Spin-polarized electron tunneling*, Phys. Rep. **238**, 173 (1994).

[75] G.-X. Miao, M. Münzenberg, and J. S. Moodera, *Tunneling path toward spintronics*, Rep. Prog. Phys. **74**, 036501 (2011).

[76] C. Caspers, M. Müller, A. X. Gray, A. M. Kaiser, A. Gloskovskii, C. S. Fadley, W. Drube, and C. M. Schneider, *Electronic structure of EuO spin filter tunnel contacts directly on silicon*, Phys. Status Solidi RRL **1**, 1 (2011).

[77] L. Plucinski, J. Minár, B. C. Sell, J. Braun, H. Ebert, C. M. Schneider, and C. S. Fadley, *Band mapping in higher-energy X-ray photoemission: Phonon effects and comparison to one-step theory*, Phys. Rev. B **78**, 035108 (2008).

[78] C. Papp, L. Plucinski, J. Minar, J. Braun, H. Ebert, C. M. Schneider, and C. S. Fadley, *Band mapping in X-ray photoelectron spectroscopy: An experimental and theoretical study of W(110) with 1.25 keV excitation*, Phys. Rev. B **84**, 045433 (2011).

[79] J. Braun, J. Minár, S. Mankovsky, V. N. Strocov, N. B. Brookes, L. Plucinski, C. M. Schneider, C. S. Fadley, and H. Ebert, *Exploring the XPS limit in soft and hard X-ray angle-resolved photoemission using a temperature-dependent one-step theory*, Phys. Rev. B **88**, 205409 (2013).

[80] S. Suga, A. Sekiyama, H. Fujiwara, Y. Nakatsu, T. Miyamachi, S. Imada, P. Baltzer, S. Niitaka, H. Takagi, K. Yoshimura, M. Yabashi, K. Tamasaku, A. Higashiya, and T. Ishikawa, *Do all nuclei recoil on photoemission in compounds?*, New J. Phys **11**, 073025 (2009).

[81] Y. Takata, Y. Kayanuma, S. Oshima, S. Tanaka, M. Yabashi, K. Tamasaku, Y. Nishino, M. Matsunami, R. Eguchi, A. Chainani, M. Oura, T. Takeuchi, Y. Senba, H. Ohashi, S. Shin, and T. Ishikawa, *Recoil Effect of Photoelectrons in the Fermi Edge of Simple Metals*, Phys. Rev. Lett. **101**, 137601 (2008).

[82] A. X. Gray, C. Papp, S. Ueda, B. Balke, Y. Yamashita, L. Plucinski, J. Minar, J. Braun, E. R. Ylvisaker, C. M. Schneider, W. E. Pickett, H. Ebert, K. Kobayashi, and C. S. Fadley, *Probing bulk electronic structure with hard X-ray angle-resolved photoemission*, Nat. Mater. **10**, 759 (2011).

[83] A. X. Gray, J. Minar, S. Ueda, P. R. Stone, Y. Yamashita, J. Fuji, J. Braun, L. Plucinski, C. M. Schneider, G. Panaccione, H. Ebert, O. D. Dubon, K. Kobayashi, and C. S. Fadley, *Bulk electronic structure of the dilute magnetic semiconductor Ga(1-x)Mn(x)As through hard X-ray angle-resolved photoemission*, Nat. Mater. **11**, 957 (2012).

[84] C. S. Fadley, and S. Nemsak, *Some future perspectives in soft- and hard- X-ray photoemission*, J. Electr. Spectroscopy **195**, 409 (2014).

[85] S.-H. Yang, B. S. Mun, N. Mannella, S.-K. Kim, J. B. Kortright, J. Underwood, F. Salmassi, E. Arenholz, A. Young, Z. Hussain, M. A. Van Hove, and C. S. Fadley, *Probing buried interfaces with soft X-ray standing wave spectroscopy: application to the Fe/Cr interface*, J. Phys. Cond. Matt. **14**, L407 (2002).

[86] J. C. Woicik, *Site-specific X-ray photoelectron spectroscopy using X-ray standing waves*, Nucl. Instr. Meth. A **547**, 227 (2005).

[87] S.-H. Yang, B. S. Mun, N. Mannella, A. Nambu, B. C. Sell, S. B. Ritchey, F. Salmassi, S. S. P. Parkin, and C. S. Fadley, *Relationship of tunnelling magnetoresistance and buried-layer densities of states as derived from standing-wave excited photoemission*, J. Phys.: Condens. Matter **18**, L259 (2006).

[88] C. S. Fadley, S.-H. Yang, B. S. Mun, J. Garcia de Abajo, in: *Solid-State Photoemission and Related Methods: Theory and Experiment*, W. Schattke and M.A. Van Hove (Eds.), (Wiley-VCH Verlag GmbH, Berlin, 2003).

[89] S.-H. Yang, B.S. Mun, and C.S. Fadley, Synchrotron Radiation News **17**, issue 3, pages 24-29 (2004).

C 6 Electron Emission and Photoemission Microscopy

Claus M. Schneider
Peter Grünberg Institut
52425 Forschungszentrum Jülich

Contents

1 Introduction **2**

2 High-Resolution Full-Field Microscopies **3**

3 Technical Aspects of Electron Emission Microscopy **4**
3.1 Electron-Optical Considerations 4
3.2 Transmission and Lateral Resolution 6
3.3 Energy-Filtered EEM 7

4 Contrast Mechanisms in EEM **9**
4.1 Primary Contrast Mechanisms 9
4.2 Secondary Contrast Mechanisms 14

5 Application to Functional Materials I: Magnetism **15**
5.1 Magnetic X-ray Circular Dichroism (MXCD) 16
5.2 Magnetic X-ray Linear Dichroism (MXLD) 22
5.3 Magnetization Dynamics Visualized in XPEEM 25

6 Application to Functional Materials II: Nonmagnetic Systems **28**
6.1 Redox Processes in Resistive Oxides 28
6.2 Overcoming the Information Depth Barrier 30
6.3 Probing the Photoelectron Spin 32

7 Concluding Remarks **35**

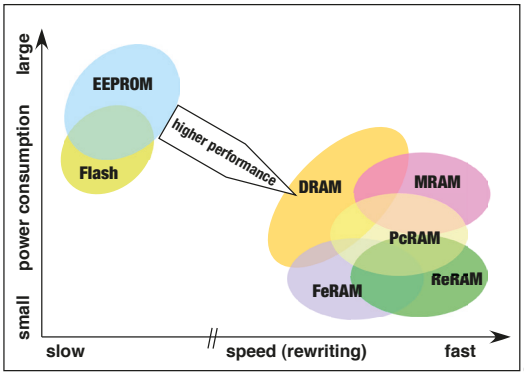


Fig. 1: *Classification of semiconductor memories (Flash, EPROM, DRAM) and alternative approaches (FeRAM, PcRAM, ReRAM, and MRAM) with respect to power consumption and rewriting speed.*

1 Introduction

Surfaces, interfaces, and nanoscale objects are crucial ingredients in modern technology. The ongoing trend for smaller and yet more powerful devices in information technology pushes the relevant lateral dimensions far into the sub-micrometer regime. Likewise, the device functionality often involves well-defined sharp interfaces, controlled concentration gradients, or even defect-like structures. In semiconductor microelectronics, for example, the smallest lateral dimension of elements in a Random Access Memory (RAM) cell is currently reaching down to about 65 nm and the 45 nm technology node is coming within reach this year [1]. A similar evolution drives magnetic data storage and spintronics. The minimal bit dimension in commercial magnetic hard disks (density of ~ 500 Gbit/in^2) has shrinked to about 15 nm [2, 3], and yet higher storage densities resulting in smaller bit sizes are demonstrated in various research labs throughout the world. At the same time, the relevant vertical dimensions have dropped into the nanometer regime. In order to observe single electron tunneling phenomena in nonmagnetic or spin-dependent transport effects in magnetic systems, extremely thin tunneling barriers of 1 - 2 nm thickness are needed. Paired with this technological progress is a strong scientific activity in surface physics, surface chemistry, and materials science, which also includes the creation of a wide variety of nanoscale systems by means of nanopatterning or self-organization processes.

The search for alternative approaches beyond Si-technology involves a wide variety of material classes. This may be illustrated for the field of nonvolatile memories (Fig. 1). In magnetic RAM (RAM) these range from ferromagnetic alloys (FeCoB) through antiferromagnets (PtMn) to insulators serving as spin-dependent tunneling barriers (MgO)[1]. Ferroelectric RAM (FeRAM) involves ferroelectric thin films such as $BaTiO_3$ or $Pb(Zr_xTi_{1-x})O_3$. Resisitive RAM (ReRAM) concepts are often based on redox processes in oxides such as Ta_2O_5 or TiO_2, or employ crystalline phase changes in chalcogenides such as GeSbTe or AgInSbTe (PcRAM). Each nonvolatile memory concept involves not only a considerable chemical complexity, but also very specific physical mechanisms and time scales which determine the read/write processes.

[1]The materials mentioned only serve as illustrative examples.

2 High-Resolution Full-Field Microscopies

The situation sketched above asks for high-resolution imaging techniques, in order to visualize the system itself and to investigate its underlying physical and chemical properties on a small lateral scale. These techniques must have a certain surface sensitivity, if studies of surface-related aspects or thin film systems are concerned. In experimental realizations, two principal imaging approaches may be distinguished. In *scanning probe techniques* a finely focused electron beam (Scanning Electron Microscopy, SEM) [4], photon beam [5], or a sharp tip with nanometer tip radius (Scanning Tunneling Microscopy, STM) [6] is scanned across the sample and the information is collected sequentially on a point-by-point basis. These scanning techniques have been constantly improved in lateral resolution and specialized with respect to the contrast mechanisms to meet various needs. Even dedicated variations for the study of magnetic systems have been developed. Among these are Scanning Electron Microscopy with Spin Polarization Analysis (SEMPA) [7], Magnetic Force Microscopy (MFM) [8], or Spin-Polarized STM (SP-STM) [9].

The *parallel* acquisition scheme is well-known from conventional optical microscopy, in which the illuminated surface area within the microscope's field of view is imaged by means of a video camera. The spatial resolution is limited by diffraction effects to $\delta_p \sim 300$ nm for visible light. As the de Broglie wavelength of electrons can be much smaller than δ_p a similar full-field imaging approach has been suggested using electrons. The image is either formed by the electrons reflected or scattered at the surface, in which case the technique is referred to as Low Energy Electron Microscopy (LEEM) [10]. If the electrons are excited in the solid and emitted from the surface, we talk about Electron Emission Microscopy. The most popular version of EEM uses photoexcitation (PEEM) at photon energies in the ultraviolet or soft x-ray regime. By exploiting specific contrast mechanisms, both techniques can also be used to study ferroic[2] phenomena at surfaces. This is particularly true for the PEEM technique, which recently became rather popular due to the increasing availability of highly brilliant synchrotron radiation from third generation storage ring facilities. The excitation with polarized soft X-rays (XPEEM) offers a unique combination of surface sensitivity, element selectivity, and magnetic contrast, as will be discussed in more detail below. In addition, the intrinsic time structure of synchrotron radiation enables time-resolved experiments down to the 10 ps regime.

In this chapter, we will first review basic aspects of EEM and PEEM starting on simple electron-optical systems and subsequently move to the high-end versions of these instruments and their application to spectroscopy. We will also discuss the various contrast mechanisms related to photoexcitation and their application to the study of different material systems. For details of the photoemission process itself the reader is referred to contribution **C5** by L. Plucinski. Extensive in-depth information on the entire field of electron emission microscopy may be found in a recent monograph by Ernst Bauer [11].

[2] In this context ferroic refers to ferromagnetic or ferroelectric phenomena in the widest sense.

3 Technical Aspects of Electron Emission Microscopy

3.1 Electron-Optical Considerations

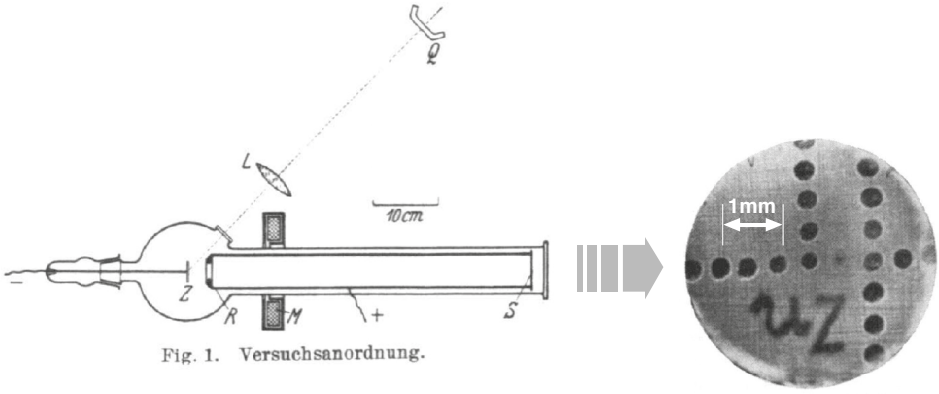


Fig. 2: *Left: Sketch of the original photoelectron microscope used by Brüche in 1934 [12]. A sample (Z) is illuminated by means of a lamp (Q) and lens (L), the emitted electrons being imaged via an aperture (R) and a magnetic lens (M) onto the screen (S). Right: Image obtained from a zinc plate with markers.*

The principal layout of an EEM exhibits strong analogies to a light-optical microscope. In order to obtain high spatial resolution the objective lens of a microscope must accept a large solid angle. In optical microscopy one therefore employs *immersion lens* objectives, which are placed very close to the object. The EEM uses a similar approach, whereby the EEM immersion lens is located $2-3$ mm away from the object surface. A first instrument of this kind has been proposed and constructed by Brüche already back in 1934 [12], employing a single magnetic lens (Fig. 2). With respect to its optical performance and large field of view, it may be thought of as a "looking glass".

Nowadays, the objective usually comprises a set of 3 (triode) to 4 (tetrode) electrostatic or magnetic ring lenses [10, 13] (Fig. 3). In contrast to the light-optical case the sample surface in an EEM forms an inherent part of the electron optical system. The immersion lens principle requires electrons to be transferred into the microscope, which leave the surface at a large starting angle relative to the surface normal. These electrons are guided by means of a strong electrostatic field ($\sim$10 kV/mm), which is applied between sample and the first electrode (extractor). This concept has two important consequences. On the one hand, it geometrically constrains the EEM experiment as the sample surface normal must be aligned with the electron optical axis in order to ensure cylindrical symmetry of the accelerating field. Non-cylindrical symmetries will cause image distortions. On the other hand, accelerating the electrons significantly reduces their relative energy $\Delta E/E_0$ and angular spread $\Delta\theta/\theta_0$, before they enter the lens system. The higher the kinetic energy E_0 the weaker the electron trajectories respond to electron optical imperfections (e.g., spherical and chromatic aberration) of the cathode lens.

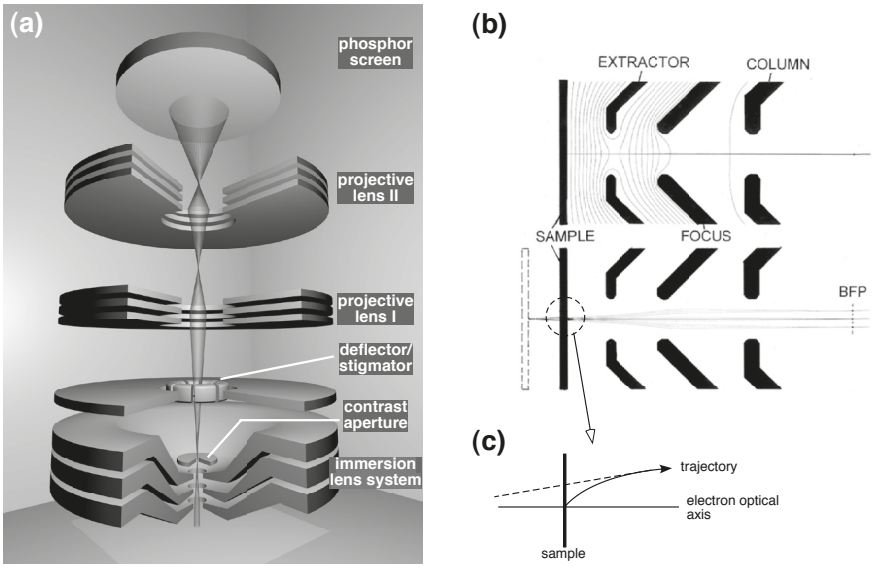


Fig. 3: *(a) Sketch of the column of an electron emission microscope. (b) Electron trajectories in a tetrode type immersion lens. (c) Trajectory in the vicinity of the surface. After [14, 15].*

After excitation in the solid, the electrons leave the surface at a starting angle ϕ_0 with a starting energy E_0. Being accelerated by the extractor field the electrons move on a curved trajectory into the objective (solid line in Fig.3(b, c)). As a result, the immersion lens, which is formed by the extractor/focus/column electrodes sees a virtual image of the sample at much lower starting angles and larger distance (Fig.3(b, c)).

For electrons starting from the same point at the surface with different kinetic energies, the lens imperfections cause the electron trajectories to fan out in the image detector plane, and will thus smear out the image point, thereby impairing the image quality and limiting the spatial resolution. In order to reduce the effect of the lens errors, we have to select electrons with the proper trajectories, i.e., a defined kinetic energy and direction. This is achieved by means of a contrast aperture, which may be located at a suitable trajectory crossover, e.g., in the back-focal plane of the objective lens. The aperture must be carefully adjusted with respect to the electron optical axis. The choice of the aperture size d is mostly dictated by practical considerations. A small aperture improves the lateral resolution, but impairs the signal-to-noise ratio and increases the image acquisition time. Practical values of d range down to 20 μm.

The remaining part of the electron optical system mainly consists of a set of projective lenses, which magnify the image onto a multichannel plate/scintillator crystal combination or a phosphor screen. This image converter transforms the “electron” into a “photon” image, which is picked up outside the vacuum chamber by a slow- or dual-scan CCD camera. An improved control of the electron beam and a (partial) compensation of electron-optical imperfections can be achieved by additional elements, such as deflector/stigmator units, which are used to bring the beam onto the electron-optical axis again. The microscope set-up is completed by a sample manipulator, which allows a lateral positioning of the sample in the field of view of the

microscope ($\sim 10 - 500$ μm, depending on microscope optics and lens settings). The small distance between immersion lens and sample limits the angle of light incidence with respect to the surface plane to $15 - 25°$.

The set-up described above contains the essential elements of a fully electrostatic PEEM. More sophisticated and elaborate electron-optical designs may also include magnetic lenses, active energy filters, and corrective elements, in order to improve the spatial resolution and spectromicroscopic capabilities of the instrument [16] (cf. Chapter 3.3).

3.2 Transmission and Lateral Resolution

The photoexcitation with soft X-ray radiation of energy $h\nu$ generates a rather broad electron spectrum $N(E_{kin})$, ranging from direct photoelectrons with maximum kinetic energy $E_{kin} = h\nu - \Phi$ (Φ : work function) down to low energy secondary electrons with ($E_{kin} \leq$ 1eV), which are a result of inelastic scattering processes in the sample. The limiting situation is that of threshold photoemission, where the photon energy is only slightly higher than the work function Φ . In this case, almost monoenergetic electrons are emitted in a narrow energy interval around the Fermi level (Fig. 4a). This popular operation mode can be conveniently realized in the laboratory by using, e.g., Hg high pressure discharge lamps.

For photon energies $h\nu \gg \Phi$ the spectrum will contain peaks due to direct interband transitions or core-level photoemission transitions (symbolized by E_1 in Fig. 4b) and a pronounced secondary electron "mountain" at low kinetic energies. The transmission function $T(E_{kin})$ of the electron optical system allows only a part of this electron spectrum to reach the image converter. In fact, the combination of immersion lens and contrast aperture in EEM has a pronounced low-pass behavior (Fig. 4c), i.e., the high energy side of the spectrum is suppressed. The width of the low energy interval transmitted depends – among others – on the diameter of the contrast aperture: the smaller the aperture diameter the smaller the energy width and the lower the image intensity. Still, the transmission of high-energy electrons is usually high enough to permit an imaging with these electrons, for example, those at E_1, provided that the instrument is equipped with an additional energy filter. Without an energy filter, the high-energy electrons just constitute a – mostly negligible – background to the image obtained with the low-energy secondary electrons.

The photoelectron energy spread is an important aspect when discussing the issue of spatial resolution. The dominant mechanism determining the spatial resolution in PEEM is the chromatic aberration of the acceleration field and the immersion lens elements. This becomes clear when inspecting Fig. 5, which shows the calculated contributions to the total resolution for the case of threshold photoemission, i.e. minimum energy spread. The extractor voltage has been chosen to 15 keV for this example. Under these conditions a lateral resolution of $\delta x \sim 10 - 20$ nm can be obtained with the present PEEM optics, corresponding to the minimum of the curve in Fig. 5. This has been proven also experimentally [17].

The larger energy spread associated with the soft x-ray excitation in XPEEM leads to energy-dependent trajectories via the chromatic aberration and results in a blurring of the image. Therefore, even for the same microscope settings, the resolution deteriorates when going from threshold photoemission to excitation with soft x-rays. In general, also increasing the acceleration

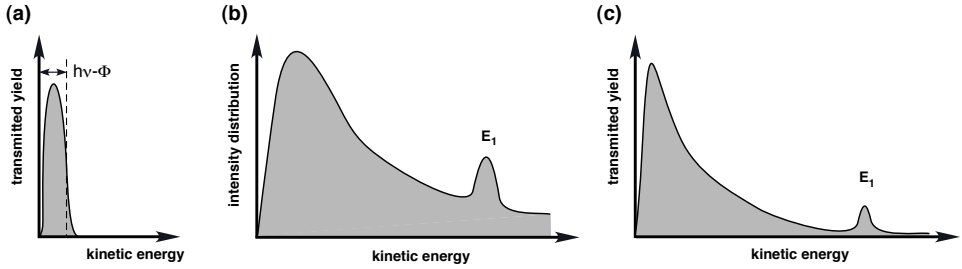


Fig. 4: *Schematic electron spectrum in the case of threshold photoemission (a). Schematic spectrum for excitation with high photon energy before (b) and after passing through the electron emission microscope optics (c). The peak E_1 marks a direct photoemission transition, e.g., from a shallow core level.*

potential and/or reducing the aperture diameter narrows down the trajectory spread at the image detector and improves the lateral resolution. The latter effect can also be clearly seen in Fig. 5. For example, a 50 μm aperture instead of a 20 μm one will provide a resolution of only 80 μm for threshold photoemission. It should be noted that for the almost monoenergetic electrons in threshold photoemission the aperture diameter can be directly translated into the maximum starting angle ϕ_0 at the surface.

There is an optimum choice of the contrast aperture diameter, however, for a given electron-optical set-up, since diffraction effects increase with decreasing aperture size. In order to stay away from the diffraction limit, the minimum aperture diameter is typically chosen as $15 - 20\mu$m. Even with the smallest aperture, the resolution achieved with soft x-ray excitation is limited to around $\delta x \leq 50$ nm [18]. A further improvement needs additional energy filtering and/or the use of aberration corrected lens systems [16].

It should be kept in mind, however, that a theoretically predicted resolution can be achieved only on an ideally flat surface, which allows an undistorted evolution of the electron trajectories. A realistic surface always has a topography with different types of defects. These range from microscopic scratches and crystalline facets down to nanoscopic features such as terrace edges and monoatomic steps. Most of these defects are associated with the formation of electrostatic microfields. These microfields superimpose with the cylindrically symmetric extractor field and cause a distortion of the trajectories. Depending on the magnitude of these microfields, the distortion leads to a trajectory spread on the image detector, which can be much larger than the surface feature itself [19]. Therefore, the lateral resolution that can be achieved with a realistic sample will be ultimately limited by its surface topography and homogeneity.

3.3 Energy-Filtered EEM

The transfer function of the EEM optics still permits the passage of high energy electrons, albeit with lower transmission. This feature can be exploited for spectroscopic imaging or *spectromicroscopy*. For this purpose, the electrons must be energy-filtered, before they are allowed to reach the image detector. There are three main approaches to select the electrons according to their energy. The first one involves electrostatic energy analyzers, which have

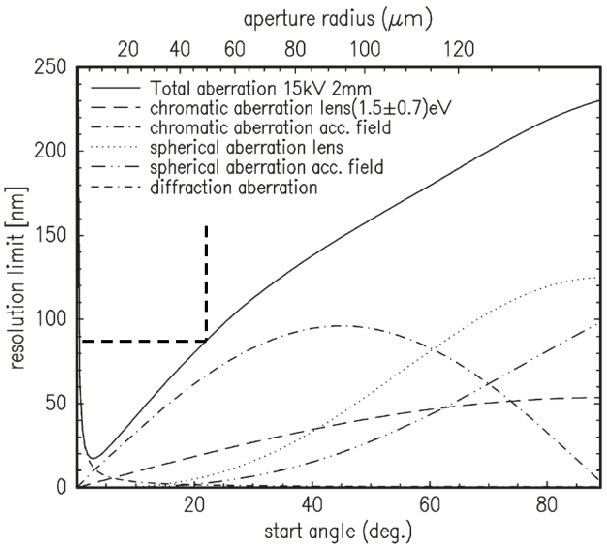


Fig. 5: *Calculated lateral resolution of a PEEM in threshold photoemission as a function of the contrast aperture size and the contributions from the various lens aberrations. From [15].*

been developed for electron spectrometers (cf. contribution **C5**).

A very common type is the hemispherical capacitor (HSC), in which the electrons are energy-dispersed within an electrostatic field build up between two concentric hemispheres (Fig. 6a). This causes a 180° turn of the electron trajectories. In order to use such a device in an EEM, however, the energy-filtering process must conserve the image. This can be achieved by proper electron-optical matching and configuring of an HSC and several of these instruments are in operation at synchrotron radiation sources worldwide. A respective layout is shown in (Fig. 6a). The hemispherical analyzer is combined with an electron-optical column based on magnetic ring lenses [20]. These have smaller spherical aberrations, but require the objective to be close to ground potential. As a consequence, a high negative voltage is applied to the sample in order to accelerate the electrons into the lens system.

An alternative approach to energy-filtering involves magnetic fields. The instrument concept in Fig. 6b describes a LEEM/PEEM system, which also uses an electron beam to illuminate the sample. In order to separate the beams travelling to and from the sample, a magnetic prism is employed, which results in a 90° bend of the electron trajectories for each pass through the prism. As a side effect, one obtains an energy dispersion of the electrons passing through the prism. By placing an aperture in the dispersive plane at the exit of the prism only electrons of a defined kinetic energy are allowed to pass and to form the image [21].

Nevertheless, the imaging quality of an analyzer is imperfect, and additional corrective actions may be needed to push the lateral resolution. A more advanced solution may thus involve two analyzers in a conjugated arrangement, in which the imaging errors of the first analyzer are compensated in the second one. Fig. 6c gives an example of such an instrument, which has been commercialized under the name of *NanoESCA* [22, 23]. Another even more sophisticated example is the SMART instrument (see constribution E4) [24], which involves an alternative conjugation concept based on an electron mirror.

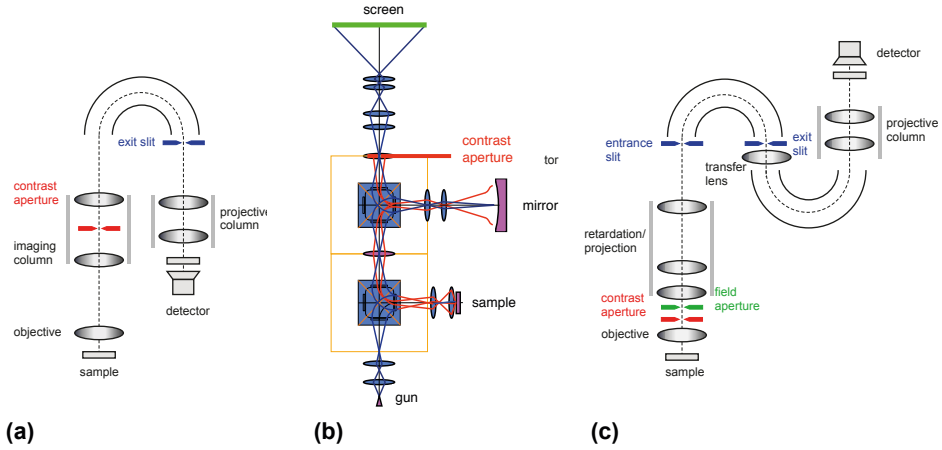


Fig. 6: *Schematic view of energy-filtered EEMs. (a) Hemispherical electrostatic energy filter (b) Energy-filtering within a magnetic prism. (c) Energy filtering with a conjugated arrangement of two hemispherical analyzers.*

4 Contrast Mechanisms in EEM

The versatility of EEM stems from the variety of physical phenomena available for image contrast formation. Formally, the image contrast can be defined as a variation of the intensity or brightness $I(x, y)$ across a field of view or an image. The contrast disappears, when $I(x, y) = const.$, i.e., the image is featureless and exhibits a uniform brightness. Therefore, the interpretation of an image requires that the origin of the image contrast be known. In the ideal case, a dominant contrast mechanism can be singled out and the image contrast can be unambigously traced back to a particular physical property that gave rise to it. For the further discussion, we will first concentrate on contrast mechanisms, which do not depend on the magnetic state of the sample. For convenience, we may also distinguish between *primary* and *secondary* nonmagnetic contrast mechanisms in the following.

4.1 Primary Contrast Mechanisms

Primary mechanisms should be understood as those processes which are intimately connected to the local electronic structure of the sample. They determine the photocurrent $I(E, \vec{k}, x, y)$ above the sample surface as a consequence of the photoexcitation step in the solid, i.e. via dipole matrix elements between the initial and final electronic states in the solid. Therefore, their origin may be related to electronic, chemical or structural aspects of the sample. For the sake of clarity, we start with the lowest excitation energies and work our way up to higher photon energies. The work function contrast widely used in threshold PEEM is the pertinent example for low-energy excitation.

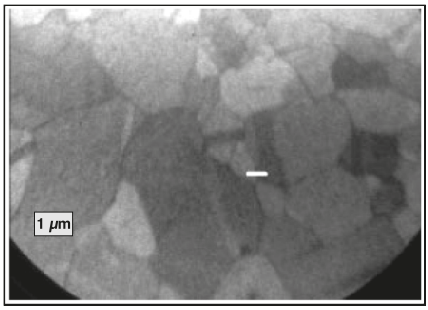


Fig. 7: *Work function related contrast on a polycrystalline Cu surface in threshold PEEM (taken from [26]).*

Work function — The work function Φ depends sensitively on the electronic state of a material and may range from $\Phi \sim 2$ eV in alkali to $\Phi \sim 5$ eV in noble metals. For a given material, it may also vary by a few 0.1 eV with the crystallographic orientation of the surface. In addition, monolayer adsorbates on a surface may also significantly affect the work function. A well-known example for the influence of adsorbates is the reduction of Φ even by submonolayer coverages of Cesium [25]. In threshold photoemission, we are able to excite electrons in the valence electronic structure, but only those excited in the vicinity of the Fermi level E_F are able to leave the crystal, causing a very narrow energy spread of the emitted photocurrent. Small changes in the work function will therefore lead to a strong modulation of the photocurrent, i.e. a chemically or crystallographically heterogeneous surface will exhibit a locally varying electron yield resulting in a contrast pattern. This way, for example, grains at a polycrystalline surface may show up in a different brightness and the contrast in the image will reflect the grain orientation [10]. A recent example for threshold PEEM from a polycrystalline Copper surface is given in Fig. 7 [26]. This contrast mechanism is dominant at low excitation energies, for example, using Hg high pressure discharge lamps, which may emit photons up to about $h\nu \sim 5$ eV. The work function contrast may be suppressed or largely masked at higher photon energies due to the relatively large energy spread of the secondary electrons contributing to the image.

Direct photoelectrons — Increasing the photon energy beyond $h\nu \sim 5$ eV, we are able to address more strongly bound electronic levels in the solid. At the same time, we are also increasing the width of the photoelectron spectrum. In an energy-filtering PEEM, a small interval of the photoelectron spectrum can be selectively imaged. A particular interesting case occurs for photon energies between $h\nu \sim 30$ eV and 100 eV, which allows the excitation of the shallow core levels in most elements. This results in sharp signatures in the photoelectron spectra, which can be exploited for an element-selective investigation [27]. A particularly instructive example is given in Fig. 8. The energy-filtered PEEM image shows a Pb film deposited onto a W(110) surface and has been acquired with the Pb $5d_{3/2}$ electrons. The film consists of a continuous monoatomic Pb layer with thicker Pb crystals (bright feature). The monolayer and the crystal have distinct electronic signatures, which are reflected in a characteristic energetic shift of the Pb $5d$ levels. By a proper setting of the kinetic energy, this difference can be exploited to yield an image contrast. We will discuss further examples for this type of approach in the subsequent chapters.

This type of contrast obtained with direct photoelectrons should no be confused, however, with the absorption-type contrast mechanism described in the following paragraph, which does not

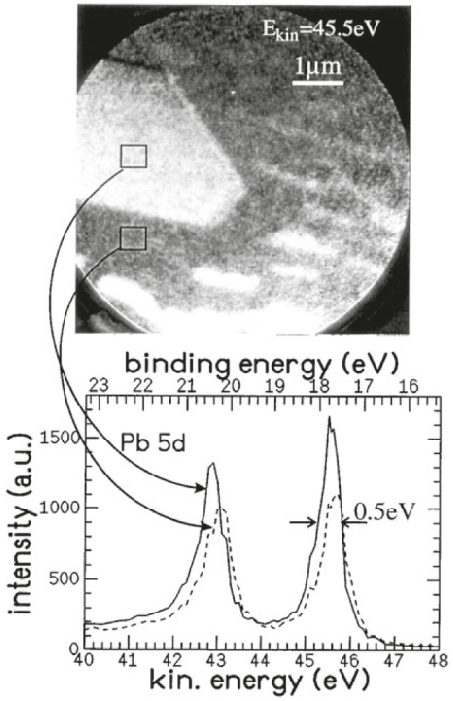


Fig. 8: *A Pb 5d photoelectron image of a Pb layer on a W(110) surface and photoelectron spectra obtained by integrating the intensity in the regions indicated. The photoelectron energy was changed in steps of 0.5-1 eV. The photon energy is 65 eV (taken from [27]).*

require an energy selection of the photoelectrons.

Secondary electrons — When working with synchrotron radiation in the soft X-ray regime, a chemical contrast may conveniently be generated by exploiting characteristic absorption edges. For this purpose, the photon energy is tuned such as to excite electrons from a core level into the empty electronic states below the vacuum level, leaving behind core holes (Fig. 9a). This process is particularly efficient in elements with only partially filled d- or f-shells, because the empty density of states is high. If the respective chemical element is inhomogeneously distributed across a surface, the absorption will be high (low) in regions where the element is present (absent). In order to utilize this absorption contrast in XPEEM the absorption signal has to be translated into electrons emitted from the sample.

A very efficient translation mechanism is provided by the electronic deexcitation of the system. The core hole created will be filled either by a radiative (X-ray fluorescence) or a non-radiative (Auger) process. In the energy regime in question, the probability for a non-radiative transition is much higher. The core hole decay leads to the emission of highly energetic Auger electrons, the total number of which (Auger electron yield) is proportional to the absorption signal (Fig. 9b). Given a proper energy discrimination, this signal can already be used to map different elements in EEM.

While passing through the solid the Auger electrons suffer multiple inelastic scattering events,

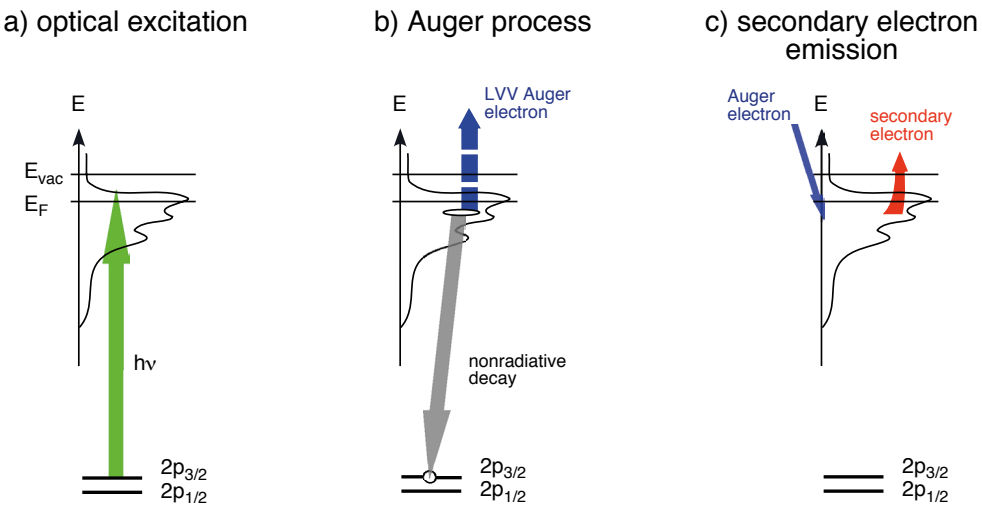


Fig. 9: *Chemical contrast mechanism in XPEEM. (a) Excitation of a core level. (b) Decay of the core hole by an Auger process. (c) Auger electron induced secondary electron cascade. From [15].*

finally leading to a secondary electron cascade. This cascade contains a large number of low energy secondary electrons, the energy distribution of which is cut off by the vacuum level (Fig. 9c). Since the starting point for this secondary electron distribution is the core hole decay and the Auger electron generation, the secondary electron yield also contains the chemical information. The low-pass characteristics of an immersion lens EEM is particularly well suited to pick up this signal.

In the simplest case, an XPEEM image of a chemically heterogeneous surface should directly map the distribution of a particular element selected by the corresponding photon energy. In other words, sample areas containing this element should appear bright in the image. This simple interpretation does not hold for every case, however, because of the various physical processes involved in the generation of the secondary electron cascade. The secondary electron yield may be strongly affected by the morphology and structure of the specimen, as well as the surface morphology and chemical state. In Fig. 10 we show an example for a Permalloy ($Fe_{20}Ni_{80}$) film on a SiO_2 surface. The film has been patterned into squares of 20×20 μm^2 size. The image (Fig. 10b) has been acquired with the photon energy tuned to the Fe L_3 edge. Naively, one would expect the Permalloy areas to appear bright under these circumstances. The opposite is obviously the case.

This observation can be explained on the basis of selected area absorption spectra (SAAS). For this purpose two alternative approaches may be used. First, a series of images is recorded as a function of photon energy (spectromicroscopy). After the series is completed, the contrast level at the selected area in each image is determined and compiled to result in a spectrum. Second, selected areas are defined by electronic means during the image acquisition, and the contrast level in these areas is measured directly while scanning the photon energy (microspectroscopy). The latter technique has been employed to obtain the SAA spectra shown in Fig. 10a. In the

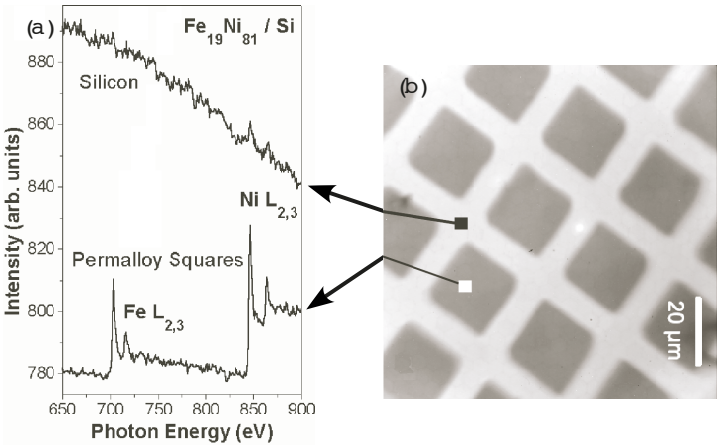


Fig. 10: *Microspectroscopy from a patterned Permalloy sample. (a) Local absorption spectra taken from the indicated areas being less than* $1\mu m \times 1\mu m$ *in size. (b) Image acquired at the Fe* L_3 *edge. From [28].*

spectrum taken on the Permalloy squares the characteristic absorption lines of Fe and Ni are easily discernible. This Permalloy signature is absent in the spectrum taken on the strip between squares, consisting of exposed SiO_2. A very weak signal at the position of the Ni lines suggests that a minute amount of Ni has been incorporated into the SiO_2 either by the ion milling process used for microstructuring or by diffusion of Ni during a short annealing step after the sample was introduced into the XPEEM chamber. At the energy position of the Fe L_3 absorption line, however, the absolute intensity level of the SiO_2 spectrum is significantly higher than the peak intensity at the Fe L_3 line of the Permalloy spectrum. This means that at this photon energy the rate of secondary electron production is higher for SiO_2 than for Permalloy. Therefore, the contrast appears "reversed" with respect to intuition. This example emphasizes two important issues. On the one hand, it demonstrates the capability of XPEEM to obtain local spectroscopic information from areas in the sub-micrometer regime. On the other hand, it illustrates that XPEEM imaging should always be accompanied by appropriate spectroscopic investigations in order to be able to unambiguously conclude on the physical or chemical origin of the image contrast.

The microspectroscopy capabilities of XPEEM can be widely employed to address materials science related issues. There often the problem arises to discriminate between different modifications of the same chemical element rather than different elements. An example for the analysis of carbon films is given in Fig. 5 [29]. Because of their hardness diamond-like carbon films are used for protecting surfaces against mechanical wear. During preparation of the films, however, also unwanted (because mechanically softer) graphitic phases may be formed. These can be spectroscopically distinguished by their higher amount of sp^2 coordinated bonds which leads to a slightly different energy position of the C-K absorption edge as compared to that observed in sp^3 coordinated diamond. This can be clearly seen by comparing microspot spectra obtained from (001) oriented diamond-like films and highly oriented pyrolythic graphite (HOPG). The predominant spectral features arise due to excitations into specific unoccupied

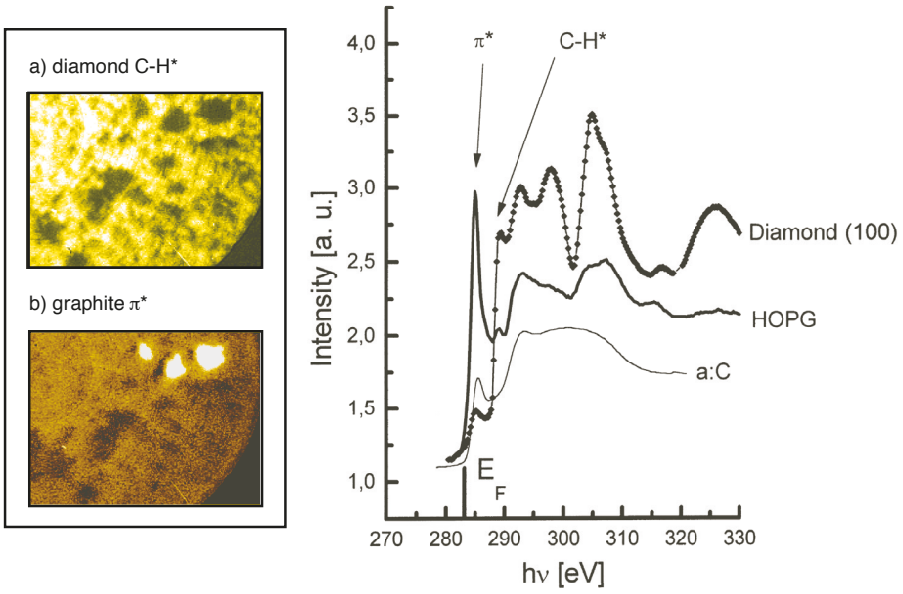


Fig. 11: *XPEEM studies of carbon films. Right: X-ray absorption spectra of a diamond (100) film and a graphite film (HOPG), both recorded in* $5 \times 5\mu m$ *microspots with XPEEM. An a-C (amorphous carbon) spectrum is shown for comparison. Left: XPEEM images from a DLC (diamond-like carbon) hard coating film recorded at photon energies corresponding to the* $C - H^*$ *and the* π^* *electronic excitations. Length of the images* $50\mu m$*. Taken from [29].*

molecular orbitals and are associated with a $1s \rightarrow \pi^*$ (HOPG) transition and the $C - H^*$ resonance (diamond). The latter is typical for sp^3 coordinated carbon with a high hydrogen content. XPEEM images recorded at these particular excitation energies reveal graphitic inclusions or contaminations in an otherwise diamond-like film (bright spots in Fig. 5b) [29].

4.2 Secondary Contrast Mechanisms

After the electrons have left the sample they experience the electrostatic field between sample surface and immersion lens. The field distribution determines the electron trajectories. Secondary contrast mechanisms refer to processes which change the electron trajectories as compared to the ideal situation. This can be achieved by local electrostatic fields (microfields) at the surface. The major source for these microfields are topographical defects, such as scratches, hillocks, or edges of geometrical structures. These create local deviations from the ideal cylindrical electrostatic field distribution in the vicinity of these defects, which in turn leads to an topographical image contrast. The way how this contrast is actually seen in the image depends strongly on the experimental parameters (shape and size of the defect, acceleration voltage, position of contrast aperture, etc.). Some examples are given in Fig. 12.

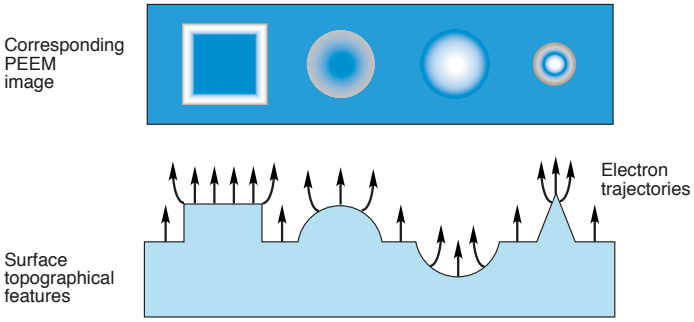


Fig. 12: *Topographic contrast patterns in PEEM arising from particular surface structures. After [30].*

Regions of different conductivity, which may lead to a partial charging up of surface areas, may also result in an image contrast. As far as magnetic samples are concerned, these are usually associated with a long-ranged magnetic stray field above the sample surface. Electrons moving in this stray field experience a Lorentz force, which will change the electron trajectories and will give rise to an magnetic image contrast. This contrast mechanism was actually employed to obtain the first PEEM images of magnetic domains [31]. The surface electrical polarization in ferroelectric materials will also affect the photoelectron emission and trajectories, leading to a distinct contrast. [ref].

It is important to keep in mind that the contrast observed in an arbitrary EEM image usually reflects a complex combination of the above contrast mechanisms. By means of appropriate experimental procedures, for example, taking the difference between two images recorded at and slightly in front of the absorption edge, or comparing images taken with different contrast aperture settings, the individual contrast contributions can be identified and extracted.

5 Application to Functional Materials I: Magnetism

The full power of the PEEM technique is unleashed if the physical phenomena giving rise to an image contrast depend on the polarization of the incident light. These may be, for example, spatially oriented electronic orbitals, which are probed by linearly polarized light. The most prominent use of PEEM in polarization dependent studies, however, concerns magnetic materials. In the following, we will therefore review *magnetic* contrast mechanisms involving polarized synchrotron radiation. This will also serve the illustrate some general properties of the XPEEM approach, which are also very useful for studies on nonmagnetic systems.

Spectroscopic studies of magnetic materials have been greatly facilitated by the discovery of X-ray magnetic dichroism with linearly (MXLD) [32] and circularly polarized light (MXCD) [33]. These effects show up in the total and partial electron yield [34], and are thus well suited as element-selective contrast mechanisms in PEEM. In addition, there are also magnetodichroic effects in threshold photoemission [35], which we have no space to cover here.

5.1 Magnetic X-ray Circular Dichroism (MXCD)

MXCD is the appropriate contrast mechanism to image *ferromagnetic* systems. It works particularly well for the transition metal $L_{2,3}$ absorption edges, because it combines a high intensity signal with a strong magneto-dichroic effect.

Physics of the magnetic contrast mechanism — The physical principle of the magnetic contrast mechanism is similar to that discussed in section 4.1. At the absorption edge, photoexcitation takes place into the unoccupied density of states (DOS) below the vacuum level. In the $3d$ transition metals, such as Fe, Co, Ni, the incompletely filled d-shell results in a high unoccupied DOS. As a result, a strong spectral feature ("white line") is observed. In addition, the ferromagnetic ground state is associated with a *spin-split* DOS (caused by the exchange interaction), which is the first important ingredient in MXCD. The $L_{2,3}$ absorption process involves the $2p$ core levels, which are spin-orbit split into $2p_{3/2}$ and $2p_{1/2}$ states by about 10 eV. This is the second important ingredient in MXCD, because photoexcitation of these states with circularly polarized (c.p.) light renders the excited electrons *spin-polarized*, even in nonmagnetic materials.

The mechanism for this spin polarization is provided through relativistic dipole selection rules [36]. In a simple picture, the angular momentum of the photon is transferred to the photoelectron and by the spin-orbit coupling only a distinct spin character is selected in the transition. The spin polarization vector $\vec{P}$ is aligned with the direction of light incidence $\vec{q}$. Depending on the light helicity $\vec{\zeta}$ ($\vec{\zeta}$ points parallel/antiparallel to $\vec{q}$ for left/right handed c.p. light) and the core level involved in the transition, $\vec{P}$ points either parallel or antiparallel to $\vec{q}$. The sign of the spin-orbit coupling is opposite for the L_2 and L_3 edges and therefore also the spin polarization $\vec{P}$ changes sign for excitation of the $2p_{3/2}$ and $2p_{1/2}$ states.

In a non-magnetic material, reversing $\vec{\zeta}$ also changes the sign of $\vec{P}$, but renders the transition probability (intensity) the same. In a magnetic material due to the spin-splitting in the empty DOS this is no longer true. Consequently, there are more empty minority than majority spin states above E_F. Therefore, if the excited core electron has a minority spin character, the transition probability will be higher than for majority spin character (Fig. 13a). The result is a magnetic dichroism, i.e., the intensity of the absorption line varies as a function of $\vec{\zeta}$ and the magnetization $\vec{M}$. The MXCD signal thans changes *both* sign and magnitude when going from the L_3 to the L_2 edge.

The translation of the MXCD signal in the photoabsorption process into an electron yield signal involves the same steps already described in section 4.1. The first step converts the MXCD into a dichroism in the Auger electron yield (Fig. 13b). Given a proper energy discrimination, this signal can already be used to obtain a magnetic image in EEM [37]. This approach provides an extreme chemical selectivity, since both the photoexcitation and the electron imaging employ characteristic spectral features. The second step finally translates the Auger MXCD signal into a helicity-dependent difference in the secondary electron yield (Fig. 13c), which may be conveniently picked up by XPEEM [38]. Detailed analyses show that the MXCD in secondary electron yield is proportional to the MXCD absorption signal [39].

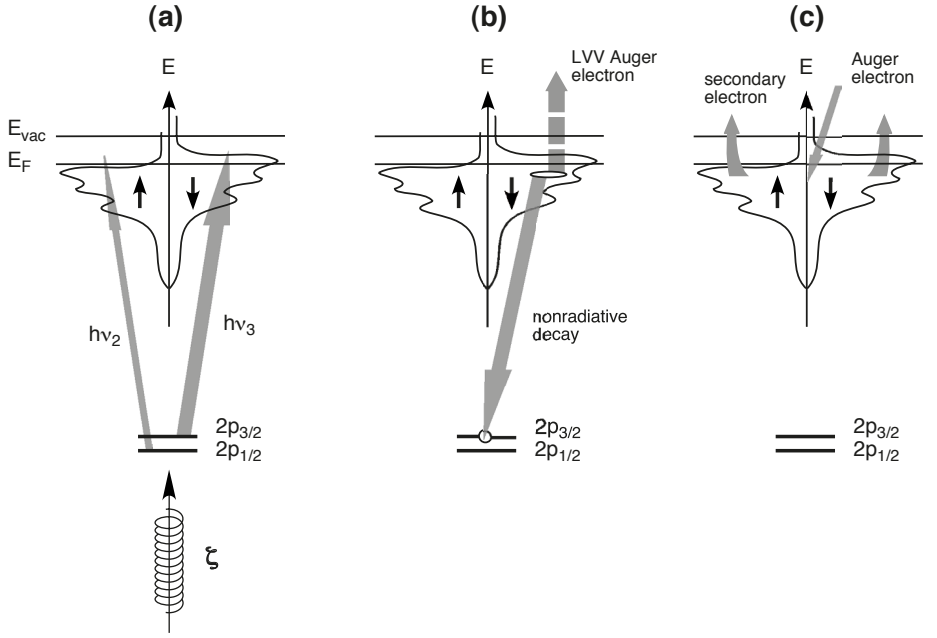


Fig. 13: *Three step process leading to magnetic contrast mechanisms based on magnetic circular dichroism. (a) Photoexcitation and generation of core holes in an $L_{2,3}$ absorption process, (b) core hole decay via Auger electron emission, (c) Auger electron induced secondary electron cascade. From [15].*

Contrast enhancement — In order to separate magnetic and nonmagnetic contributions to the contrast in the EEM image, one conveniently uses the fact that a reversal of $\vec{\zeta}$ changes the sign of the magneto-dichroic signal C_M, while leaving the nonmagnetic signal C_{NM} unaffected, i.e.,

$$C_M(-\vec{\zeta}) = -C_M(\vec{\zeta}) \;\; ; \;\; C_{NM}(-\vec{\zeta}) = C_{NM}(\vec{\zeta}) \;\; . \tag{1}$$

Therefore, by subtracting two images taken at the same photon energy, but opposite light helicity I_{ζ^+}, I_{ζ^-}, the magnetic contrast C_M is enhanced. Summing up the two images extracts the non-magnetic contrast C_{NM}

$$C_M \sim I_{\zeta^+} - I_{\zeta^-} \;\; ; \;\; C_{NM} \sim I_{\zeta^+} + I_{\zeta^-} \;\; . \tag{2}$$

This way, the images provide both magnetic and nonmagnetic (chemical, topographical) information. In order to describe the magnetic contrast in a more quantitative manner, often the *asymmetry* image A

$$A = \frac{I_{\zeta^+} - I_{\zeta^-}}{I_{\zeta^+} + I_{\zeta^-}} \;\; . \tag{3}$$

rather than the difference image C_M is shown. The quantity dichroic asymmetry ranges between $+100\%$ and -100%.

An example for this contrast enhancement procedure is shown in Fig. 14. The sample consisted of Permalloy micropatterns and the images have been recorded at the Ni L_3 edge. The

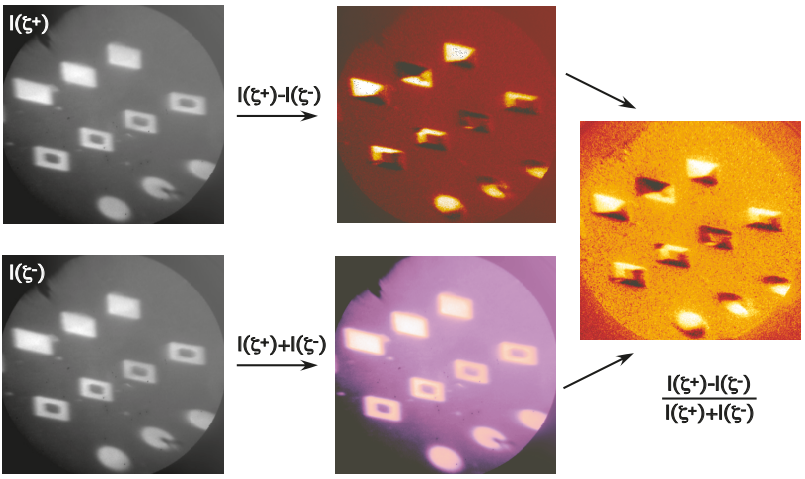


Fig. 14: *Magnetic contrast enhancement for the example of Permalloy microstructures. Individual images (left) taken with opposite light helicities $I(\zeta^+)$ and $I(\zeta^-)$ at the Ni L_3 edge are first subtracted and summed up (center), respectively. The sum and difference images are finally used to calculate an asymmetry image (right). Feature size is 12μm. See also [15].*

magnetic contrast is rather weak and does therefore not show up directly in the images taken at opposite helicity $I(\zeta^+)$ and $I(\zeta^-)$ – the patterns appear in an almost uniform gray level. The same is found for the sum image in which the magnetic information should be eliminated. The difference image, however, reveals a clear internal structure in each micropattern, reflecting the lateral distribution of magnetic domains. The dark and bright areas correspond to magnetic domains with different spatial orientations of the magnetization vector (see below). The asymmetry image shows the same magnetic contrast on a normalized scale. Whether difference or asymmetry images are used for the analysis of problems in surface and thin film magnetism will depend on the actual experimental situation.

Angular dependence of the image contrast — The magnetic domain images obtained by XPEEM contain also quantitative information on the local orientation and distribution of the magnetization vector $\vec{M}(x, y)$. The geometrical relationship between magnetization vector $\vec{M}$ and light helicity $\vec{\zeta}$ results in the magnitude of the magnetic contrast scaling as

$$A \sim \vec{M} \cdot \vec{\zeta} \ . \tag{4}$$

This behavior is nicely illustrated by XPEEM imaging of magnetic domains at single crystal surfaces. The image in Fig. 15 has been acquired exploiting MXCD at the Fe $L_{2,3}$ edges. The magnetic contrast was enhanced using the procedure given in eq. (2).

Fig. 15 shows the example of a four-fold symmetric surface – an Fe(001) whisker. Iron whiskers are known to exhibit very large regular domains. In the image, we can discern four different contrast levels (black, white, dark gray, light gray). This has been achieved by rotating the direction of light incidence by about $\phi = 15°$ away from the in-plane $\langle 100 \rangle$ direction. This

leaves each easy axis of magnetization with a non-zero projection along $\vec{\zeta}$ (projection angles $15°$, $75°$, $105°$, $165°$). According to eq. (4) this will result in a distinct contrast level for each magnetization direction. This is indeed found in the experiment, when looking at the statistic distribution of contrast (gray) levels in the image. The 8-bit representation of the image results in 256 gray levels, with the histogram revealing four clear broad maxima corresponding to the domain orientations in the image.

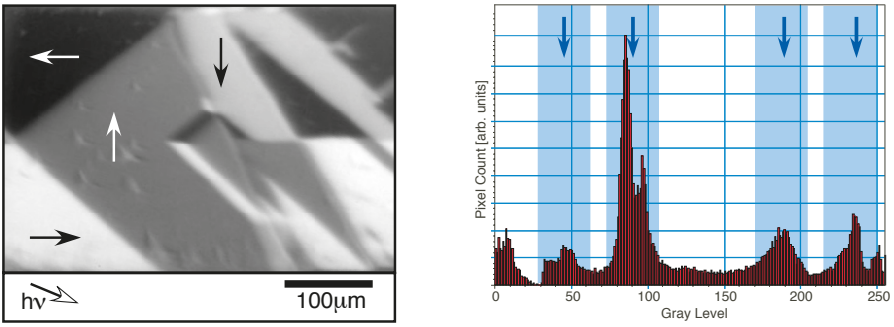


Fig. 15: *Magnetic domain pattern on a ferromagnetic Fe(001) whisker surface obtained at the Fe $L_{2,3}$ edge (arrows mark the local orientation of the magnetization vector). Right: Histogram of the gray levels in the image revealing four broad maxima (shaded regions) corresponding to distinct magnetization directions. From [15].*

With respect to details of the magnetic microstructure, in the left-hand side of the picture a classical flux closure pattern has formed, whereby neigboring domains are bounded by $90°$ domain walls. A more complicated situation is found on the right-hand side of the image, where smaller domains appear. This complex domain pattern indicates that the whisker is actually not in an ideal state, and the magnetic microstructure is largely determined by the mechanical strain in the system. An annealing procedure brings the whisker into a strain-free state, which is characterized by a very simple domain pattern (see Sect. 5.1).

Magnetic domain walls — Mesoscopic spin structures which pose a considerable challenge to magnetic imaging experiments are the boundaries of magnetic domains, the domain walls. Given two domains with local orientation of the magnetization $\vec{m}_1$ and $\vec{m}_2$ separated by a domain wall, the magnetization $\vec{M}$ must continuously rotate from $\vec{m}_1$ to $\vec{m}_2$ within the width of the domain wall. This can be done in two principal ways [40]. In a *Bloch* type domain wall $\vec{M}$ rotates within the plane of the wall, i.e., the component of $\vec{M}$ normal to the wall plane is always zero. By contrast, in a *Néel* type domain wall $\vec{M}$ rotates in a plane perpendicular to the wall plane, and thus always has a perpendicular component perpendicular to the wall.

The angular dependence of MXCD (eq. 4) can be utilized to selectively image magnetic domain walls under certain circumstances [41]. The example shown in Fig. 16 has been obtained from a thick Fe(001) whisker, which had developed a peculiar domain structure. Since the light was incident along an easy axis of magnetization, the pattern consisting of large domains (Fig. 16a) is again characterized by three distinct contrast levels. These are associated with $\vec{M}$ parallel, antiparallel, and orthogonal to the incoming light ($\vec{\zeta}$). A closer inspection of the image, however,

reveals two narrow straight lines in the center of the pattern bounding a diamond shaped area. These lines start at the boundaries of the left-hand (black) and right-hand (white) domain. Each line has a distinct contrast level, which changes only at the point where they meet. These findings suggest that the lines in the image are caused by domain walls.

The origin of such a domain wall contrast can be understood, if the actual surface termination of a domain wall is taken into account. It is known for Fe(001) that Bloch walls in the bulk form a Néel like surface termination [42]. This is due to an energy (stray field) minimization argument. A Bloch-like rotation of $\vec{M}$ would lead to a magnetization component perpendicular to the surface. In order to avoid this energetically unfavorable situation, the Bloch wall continuously transforms into a Néel like wall when it reaches the near-surface region [43]. As a result, the Néel type rotation of $\vec{M}$ takes place within the surface plane. If we take the example of two domains with opposite direction of $\vec{M}$ ($180°$ domain wall), the magnetization vector must rotate within the surface by $180°$. Recalling the angular variation of the MXCD signal (eq. 4) somewhere during this rotation the $\vec{M}$ will have a sizable component along $\vec{\zeta}$, giving rise to distinct contrast. It should also be noted that the rotation of $\vec{M}$ across the $180°$ domain wall can take place with either a right- or a left-hand turn, since these two cases are energetically equivalent. Consequently, one should expect *two* distinct contrast levels for a domain wall, just as is observed in the experiment (Fig. 16).

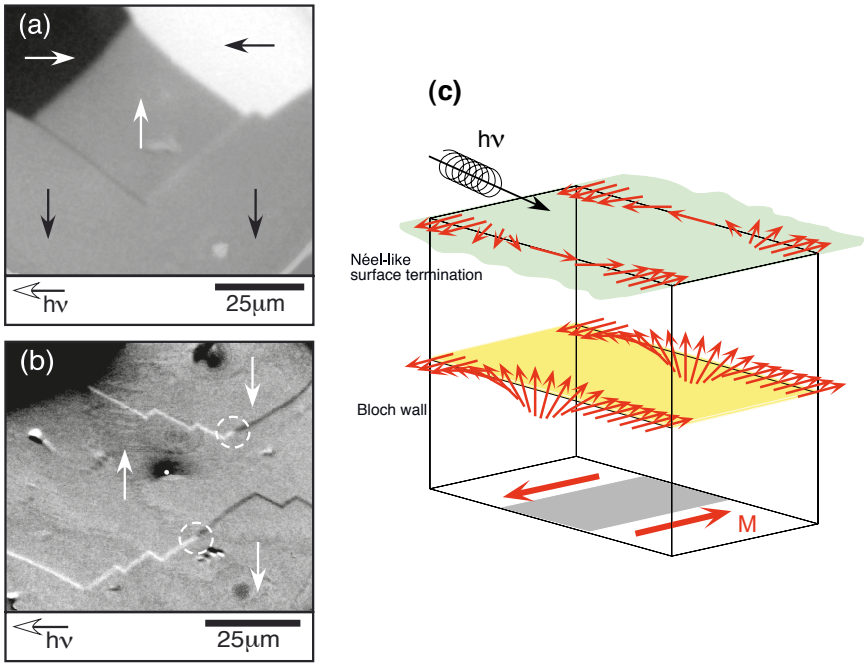


Fig. 16: *Magnetic domain pattern at the surface of a strained Fe(001) whisker. (a) Domain and domain wall contrast; (b) Selective imaging of domain walls (white and dark lines). After [41].*

The above interpretation can qualitatively explain the experimental findings. The actual situation, however, is more complex than a simple $180°$ wall. This can be easily seen by recon-

structing the magnetization distribution in Fig. 16a. The domain pattern is energetically very unfavorable, because the head-on orientation of the local magnetization directions generates a magnetic stray field. Such a domain structure cannot be explained by surface effects solely. In fact, the pattern in Fig. 16a must be interpreted on the basis of closure domains stabilized by the magnetic microstructure in the bulk, with the magnetization direction of the bulk domains pointing perpendicular to the surface [44]. In this case, the domain walls can be identified as socalled "V"-lines [45], being a result of two 90° bulk domain walls meeting at the surface. The formation of V-lines is caused by mechanical strain in the crystal.

A characteristic property of V-lines is a "zig-zag" course of the domain walls, which can be seen in Fig. 16b. In this image the main magnetic contrast arises from the domain walls only. The angle between two neighboring segments of the zig-zag line has previously been determined to about 109° [45], which is compatible with the result in Fig. 16b. Finally, we also observe the predicted jump in the contrast level along the course of a domain wall (encircled regions), associated with a change of the rotation sense of $\vec{M}$ at the surface.

Information depth — An important aspect of the contrast mechanism is the magnetic depth of information. In the case of MXD effects, it is determined by two factors: *(i)* inelastic mean free path of the electrons λ_{in}, and penetration depth of the incident light λ_p. Although the low energy secondary electrons are the same imaged in a SEMPA experiment, there is a fundamental difference. SEMPA determines the actual *spin polarization* of these electrons, which is determined by spin dependent scattering processes in the magnetic material and limits the information depth to about $5\mathring{A}$ [46]. In XPEEM, however, the *intensity* of the low energy electron cascade (yield) is measured. The information depth is thus determined by the escape depth of the Auger electrons and the physics of the cascade formation process. It reaches values of the order of $\lambda_{in} \approx 15 - 25\mathring{A}$ in ferromagnetic metals [39]. The penetration depth of the incident light is usually significantly larger than λ_{in}. If λ_{in} becomes comparable to the electron escape depth, for example, due to strong absorption or grazing incidence, then considerable saturation effects may occur in the MXD signals [47, 39]. As in a typical XPEEM geometry the light impinges at an angle of 15 - 25° with respect to the surface, saturation effects may have to be considered in quantitative measurements.

The following example illustrates the combination of chemical and magnetic information which can be extracted from XPEEM investigations. The sample consisted of an epitaxially grown Cr wedge (ranging from 0-4 monolayers (ML)) on a Fe(001) whisker surface. The Cr wedge has then been covered by a 5 monolayer Co film. The magnetic domain pattern in this sample has then been imaged in the "light" of the Fe, Co, and Cr L_3 absorption lines. Due to the information depth of the MXCD approach discussed above, the domain patterns of the substrate and the Cr layer can be imaged through the respective overlayers. The separation into magnetic and chemical information follows from (eq. 2).

The results are compiled in Fig. 17. The whisker surface has a particularly simple domain pattern, consisting of only two oppositely magnetized domains. This corresponds to the equilibrium domain structure of a strain-free iron whisker. The chemical information about the Cr wedge confirms the onset (marked by the broken line, due to the graphical reproduction the first part of the wedge may appear dark) and a subsequent linear increase of the Cr signal along the wedge direction. Because of technical circumstances the thickness gradient of the wedge

is inclined to the whisker main axis. The Co overlayer also shows a clear magnetic signal and a well-defined domain structure. In contrast to the Fe domain pattern, however, the Co domain structure reveals a characteristic change at a critical Cr thickness of about 2 ML. Below this critical thickness, the magnetization direction is the same in both Fe substrate and Co overlayer, i.e., the Co overlayer couples parallel ("ferromagnetically") to the substrate magnetization. Above 2 ML Cr the Co layer reverses its magnetization with respect to the substrate, i.e. it couples antiparallel ("antiferromagnetically"). This behavior is caused by the interlayer exchange coupling through the Cr wedge [48]. Depending on the thickness of the Cr film, the coupling changes its character from parallel to antiparallel and vice versa.

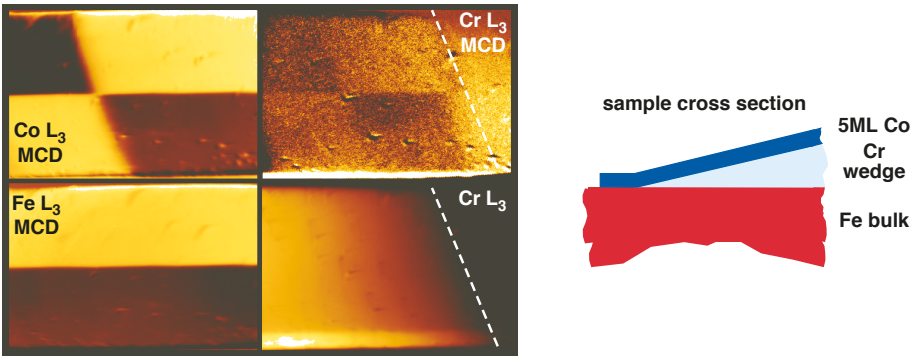


Fig. 17: *Exchange coupled thin film system Co/Cr/Fe(001) whisker (broken line marks onset of Cr wedge). Left: Compilation of magnetic domain images acquired at the Co, Fe, and Cr L_3 edges. Right: Sample cross section. Taken from [49].*

A surprising result is found in the Cr magnetic signal, demonstrating the high element selectivity and magnetic sensitivity of the XPEEM approach. First, also on the Cr L_3 edge a magnetic domain pattern is observed, even in the Cr submonolayer regime. Second, the Cr domain pattern follows closely the Co one, as the change in magnetization direction at about 2 ML Cr is easily discernible. Compared to the magnetic contrast of Fe and Co, the Cr magnetic signal is rather weak. This is consistent with earlier findings, and can be explained by a small magnetic moment of the Cr (either due to magnetic frustration or partially antiferromagnetic order in the Cr patches) [50]. The Cr signal arises due to the exchange coupling to the neighboring Fe and Co layers, causing a partial polarization of the Cr. The results suggest, however, that the coupling Co-Cr seems to be stronger than the respective Fe-Cr coupling. Below 2 ML Cr we cannot distinguish between these two coupling contributions, as the magnetization directions in Fe and Co are the same. Above 2 ML Cr the Cr signal clearly follows the Co magnetic orientation with a gradual reduction of the Cr MXCD contrast levels. The latter is related to a dilution effect caused by the unpolarized Cr atoms in the bulk of the Cr interlayer.

5.2 Magnetic X-ray Linear Dichroism (MXLD)

Another important class of materials for magneto-electronic applications are *antiferromagnets*. However, the expectation value of the local magnetization in antiferromagnets vanishes, i.e. $\langle \vec{M} \rangle = 0$. Therefore, MXCD cannot be used to image the magnetic domain structure in these

materials. Magnetic linear dichroism poses a solution to this problem, since the MXLD signal depends on $\langle M^2 \rangle$ [51].

Properties of the contrast mechanism — Antiferromagnets may have a very complex spin arrangements. Simple cases are found in Cr or NiO, in which neighboring lattice planes have opposite spin alignment. This "topological" antiferromagnetism leads to socalled uncompensated planes ($\{001\}$ in Cr, $\{111\}$ in NiO). The orientation of the magnetic moments along a spatial direction defines a quantization axis $\vec{m}$, which is used as reference for the optical excitation with linearly polarized light (π). MXLD appears between absorption spectra taken with linear polarization parallel $\pi_\parallel$ and perpendicular $\pi_\perp$ to $\vec{m}$. Therefore, the antiferromagnetic contrast C_{AFM} is extracted in analogy to eq. (2) as

$$C_{AFM} \sim I_{\pi_\parallel} - I_{\pi_\perp} \ . \tag{5}$$

The magnetic contrast is quantified by normalizing C_{AFM} to the sum of $I_{\pi_\parallel}$ and $I_{\pi_\perp}$. The angular dependence between the direction of light polarization and the quantization axis $\vec{m}$ is slightly more complex than in the MXCD case. If we keep the spin-quantization axis fixed and vary the linear polarization, the XMLD contrast behaves as

$$C_{AFM} \sim (3cos^2\theta - 1) \ . \tag{6}$$

where θ is the angle between polarization and spin-axis. Using this approach, domains at the surface of antiferromagnetic materials have been successfully imaged [52, 53].

Imaging domains in antiferromagnets — The last issue addresses the imaging of domains in an antiferromagnetic material on the basis of magnetic linear dichroism. Pioneering experiments in this field have been performed by Spanke *et al.* [52] and Stöhr *et al.* [53] on NiO films. Subsequent studies have concentrated on the microscopic mechanisms of the exchange anisotropy between ferro- and antiferromagnets [54]. The example reproduced in Fig. 18 shows the antiferromagnetic domain pattern on the (001) surface of a NiO single crystal [55]).

The domain image is the result of a specific contrast enhancement procedure. The data have been recorded with s-polarized light at the Ni L_2 edge. In NiO the absorption at the L_2 edge involves a characteristic doublet structure which causes the absorption line to consist of two spectral features ($h\nu_1$, $h\nu_2$) separated by about 1 eV [56]. These features exhibit a pronounced magnetic linear dichroism. The magnetic domain contrast is thus enhanced by calculating the asymmetry distribution (eq. 3 from two images acquired at $h\nu_1$ and $h\nu_2$, respectively.

The domain pattern itself is quite complex. The reason for this is the antiferromagnetic spin arrangement in crystalline NiO. The main "stacking" axes $\vec{m}$ are given by the four equivalent [111] directions, along which lattice planes with in-plane ferromagnetic order are stacked in an antiferromagnetic arrangement. Within a (111) plane, there are three equivalent [211] easy directions into which the spins can point. As a consequence, in the bulk there is a total of 12 different types of antiferromagnetic domains [57]. These must be projected onto the (001) surface plane in order to obtain the possible bulk-truncated surface configurations (neglecting surface-induced changes of the spin structure for the moment). Due to the angular dependence of the MXLD contrast (eq. 6) the XPEEM picks up the component of the local antiferromagnetic

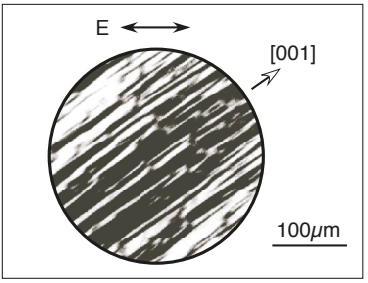


Fig. 18: *Antiferromagnetic domains in NiO observed using MXLD [55]. Direction of the light polarization ($\vec{E}$) with respect to the in-plane crystalline orientation is indicated.*

orientation vector $\vec{m}$ along the electric field vector $\vec{E}$ of the linearly polarized light. In view of the complexity of the NiO(001) situation, a series of images as a function of the angle of light incidence are needed to unambiguously reconstruct the details of the surface domain pattern.

Magnetically coupled systems — The versatility of PEEM with respect to magnetic contrast mechanisms permits a very detailed view on heterogeneous magnetic systems and magnetic coupling phenomena. As an example, we discuss a system consisting of a thin film of NiO – an antiferromagnet – on top of a Fe_3O_4 (magnetite) bulk single crystal surface – a ferrimagnet. The MXCD contrast at the Fe L_3 edge reveals the magnetic domains at the magnetite interface underneath the NiO film (Fig. 19a). The image represents the ratio between two images taken at opposite light helicity in order to emphasize the magnetic contrast. This domain pattern at the surface is determined by the configuration of the underlying bulk domains. Essentially four different contrast levels can be discerned, corresponding to four in-plane magnetization directions at the surface (marked by the single-headed arrows). In the larger part of the sample the domains are bounded by straight walls, except for the upper right region, where curved domain boundaries and smaller domains appear, presumably caused by defect-induced strain in the sample.

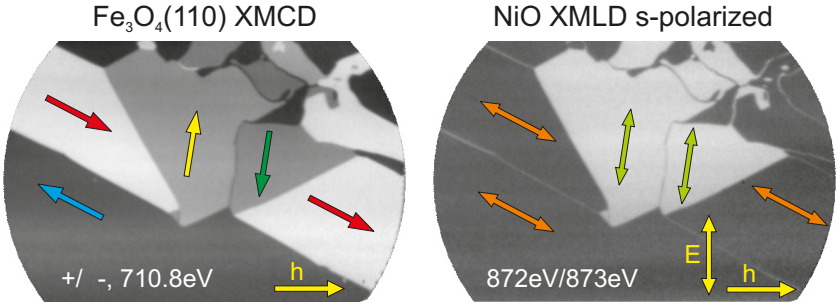


Fig. 19: *Magnetic domains in $NiO/Fe_3O_4(011)$, an antiferromagnet/ferrimagnet hybrid model system. The imaging with circularly polarized light of opposite helicity (+, -) at the Fe L_3 edge yields the ferromagnetic domains in $Fe_3O_4(011)$, whereas the illumination with linearly polarized light at the Ni L_2 edge yields the domains in the antiferromagnetic NiO overlayer. h and E denote the directions of light incidence and electric field vector, respectively.*

If we now switch to the Ni L_2 edge and take images with linearly polarized light, we will see

the domains in the NiO film. As the MXLD contrast is largest for a spectral substructure of the L_2 absorption line, the contrast enhancement procedure differs from the MXCD case in that the ratio of two images taken at slightly different photon energies $h\nu = 872$ eV and $h\nu = 873$ eV is formed. The resulting domain picture reveals the same geometrical pattern (Fig. 19), but only two different contrast levels – dark and bright. Apparently the ferrimagnet (FM) imprints its surface domain structure into the antiferromagnetic (AFM) film via an exchange coupling mechanism, which is also a crucial ingredient for exchange biasing [58]. The fact that we only find two contrast levels in the NiO film is, of course, a consequence of the antiferromagnetic spin arrangement in the NiO, which defines an axis for spin *alignment*, but not a spin *orientation* as in the ferromagnet. From a quantitative analysis of the contrast levels one finds that the local spin quantization axis in the NiO (double-headed arrows) points always along the magnetization direction in the ferrimagnet. This type of "spin-flip" coupling between ferromagnets and antiferromagnets is believed to appear as a consequence of a microscopic interfacial roughness [59].

It should be pointed out that although the magnetic contrast in the NiO may be the same for two domains, the detailed spin arrangement in these domains may differ by $180°$. This can be directly seen by comparing the ferro- and antiferromagnetic domain patterns in (Fig. 19). For domains in the underlying Fe_3O_4 being oriented into opposite directions, the antiferromagnetic domains in NiO yield the same contrast. At the same time, the exchange coupling requires the spins directly at the interface FM/AFM to be aligned parallel, which has been proven by respective MXCD studies. Assuming the antiferromagnetic layer sequence in the film to be determined by the interfacial NiO layer requires this layer sequence to differ between adjacent domains. As a consequence an antiferromagnetic domain wall should be formed between these domains. Indeed, a closer inspection of the MXLD images reveals a narrow bright line separating these two domains. This indicates that the antiferromagnetic spin configuration in the "bulk" of the NiO film must change between the two domains, although the spin alignment axis is the same.

5.3 Magnetization Dynamics Visualized in XPEEM

The parallel imaging capabilities of EEM also provide a very interesting pathway to time-resolved imaging, which can be particularly well demonstrated in magnetic systems. In order to describe the main principles, we will concentrate on the time-resolved PEEM in the following, which is presently the most advanced time-resolved EEM technique. In general, one may distinguish two approaches. In the one shot imaging approach, only a single image is acquired, capturing a certain state of the system in time. This permits an imaging of statistic events or fluctuations. In order to obtain a sufficient signal-to-noise ratio, however, one shot imaging requires extremely bright light sources, which are actually not widely available up to date. They may become available in the future in form of the free-electron lasers, which are developed at several places world-wide. An alternative is offered by the second approach, the stroboscopic or pump-probe imaging. In this case, the sample must be repeatedly excited by a signal of a well-defined time structure. In this case, the sample cannot simply be imaged, but also needs to be excited in a well-defined pulse-wise manner and time-sequence. The excitation can be performed, for example, by a short laser pulse or a current or magnetic field pulse in case of a magnetic system. After the excitation pulse the state of the sample is imaged with some time

delay. This procedure is repeated until a sufficient image quality has been achieved. In order to obtaine meaningful data, the sample must return to the same initial state after each excitation pulse and the excitation process must be absolutely reproducible. As a consequence, only reversible processes are accessible by this pump-probe procedure.

A necessary ingredient in the pump-probe experiment is a pulsed light source. The synchrotron radiation generated in a storage ring has an intrinsic time structure. This is due to the fact that the electrons circulating in the ring do not form a continuous beam, but are grouped in evenly spaced packets or "bunches" [60]. The synchrotron radiation emitted into a beamline consists therefore of light pulses with typically a few 10 ps width and a spacing ranging from a few ns to almost a μs, depending on the operational characteristics of the storage ring. This feature was exploited in first time-resolving XPEEM studies of magnetic structures [61, 62]. The pulse width of the synchrotron radiation allows a convenient access to processes taking place on time scales down to 10 ps, i.e. 100 GHz. In magnetism, this is a very interesting regime, as it coincides with the precessional frequencies in a ferromagnetic system, which in turn are relevant for the speed of magnetization reversal processes. Therefore, combining high lateral and time resolution with high element selectivity and magnetic contrast, time-resolved XPEEM (TR-XPEEM) represents an extremely powerful approach for the investigations of magneto- and spin dynamics at surfaces and in thin films.

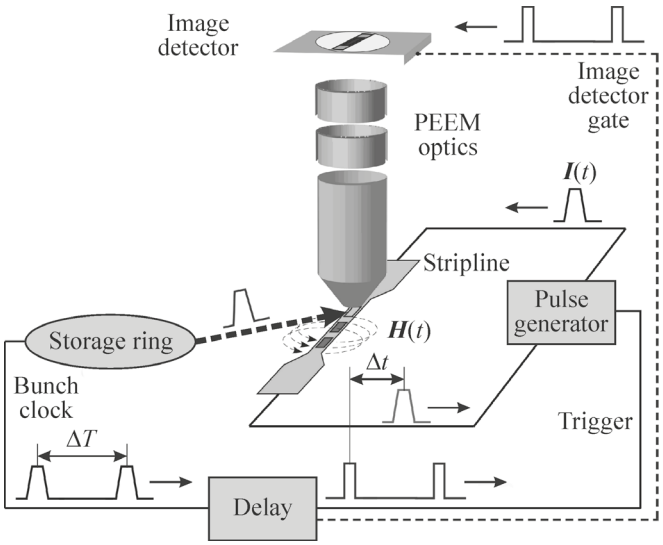


Fig. 20: *Scheme of a time-resolved XPEEM experiment from a magnetic system. [55].*

A stroboscopic imaging procedure for a magnetodynamic experiment works as follows. The system is "pumped" by a magnetic field at time t_0 and probed by the light pulse at a later time t_1, an image being taken with each light pulse. The time delay $\Delta t = t_1 - t_0$ is usually adjusted electronically. In order to obtain sufficiently fast (GHz) and strong magnetic field changes at the sample, coplanar waveguide geometries or microcoils are employed, depending whether the magnetic field vector $\vec{H}$ should be directed parallel or perpendicular to the sample surface [63]. Second, the confinement of the magnetic field in these geometries asks for small thin film

elements as samples. The magnetic field pulses can be conveniently generated by fast electrical pulse generators or photoconductive switches.

A good signal-to-noise ratio in the image is usually obtained after accumulating over $10^8 - 10^9$ pulse cycles. In the simplest case, the accumulation is performed by on-chip integration within a slow-scan CCD camera. More sophisticated designs involve gated detector schemes to enable a flexible pulse picking [63]. As mentioned above, these pump-probe experiments can only capture the *reversible* processes launched by this field pulse. Slow irreversible processes will lead to metastable states, fast ones to a washing-out of the image features.

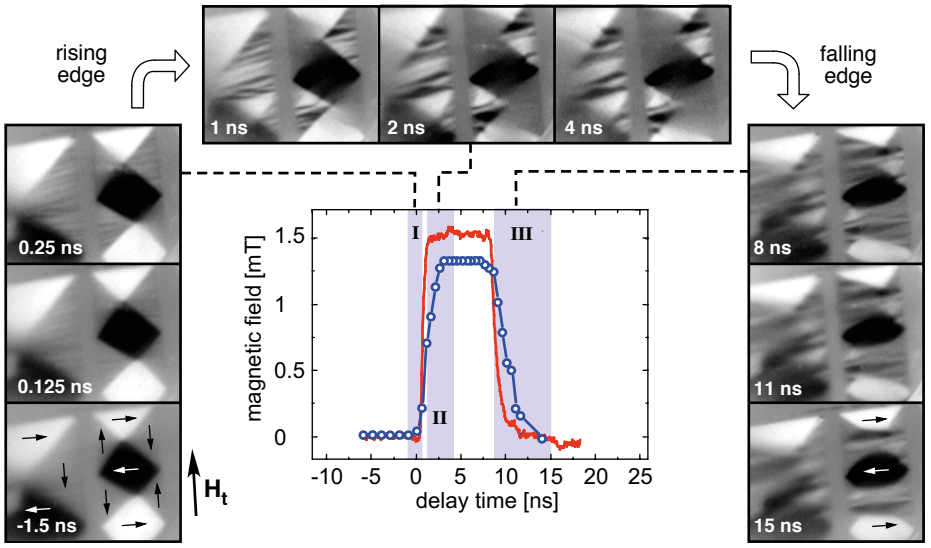


Fig. 21: *Time-resolved imaging of the magnetodynamic response of a bar-shaped Permalloy element to a nanosecond magnetic field pulse.*

An example for reversible processes is shown in Fig. 21, giving the magnetodynamic response of rectangular Permalloy microelements to a 10 ns long magnetic field pulse with a rise time of about 500 ps. The effective time-resolution in the experiment is about 50 ps. Concentrating on the top element, we note that prior to the onset of the field pulse $\vec{H}$ ($t = -1.5$ ns) the domain structure forms a standard Landau flux-closure pattern with the typical triangular and diamond-shaped domains, connected by $90°$-walls and vortices. The dark and white areas correspond to a large magnetization component antiparallel and parallel to the incoming light. In the grey areas the magnetization vector $\vec{M}$ points perpendicular to the incoming light. This image has already been recorded in the stroboscopic mode, proving that the element relaxed into this state after each field pulse. During the onset of the field pulse fine ripple-like networks of darker and brighter stripes form in those triangular domains with $\vec{M}$ antiparallel to $\vec{H}$. Note that initially the field pulse does not excert a torque onto the domain magnetization, as $\vec{M} \uparrow\downarrow \vec{H}$. These stripe networks are indicative of incoherent rotation processes [64], which locally turn $\vec{M}$ *perpendicular* to $\vec{H}$.

The magnetic contrast of the stripe networks becomes stronger as the magnetic field increases.

This is caused by a further rotation of the local magnetization towards the line of light incidence. In addition, smaller stripes coalesce into larger ones. After reaching the pulse plateau ($t = 2\,\mathrm{ns}$), the stripe pattern becomes stationary. The system assumes a new temporary equilibrium, determined by the magnetic anisotropy and the external field $\vec{H}$. If $\vec{H}$ is reduced again, this can also be seen as an effective field pointing in the opposite direction, i.e., $-\vec{H}'$. As a consequence, the stripes spread immediately also into the opposite triangular domains with $\vec{M} \uparrow\uparrow \vec{H}$ [65]. Similar observations have also been made on ring-shaped structures [66]. These stripe domain patterns are associated with a large transient stray field at the sample edge, the implications of which are discussed in Sect. 4.

The dark and bright domains at $t = -1.5\,\mathrm{ns}$ behave differently during the field pulse, as they initially experience a strong torque $\vec{M} \times \vec{H}$, resulting in a coherent rotation of $\vec{M}$ in these domains. The rotation is not homogeneous throughout the domain, as can be seen by the gradual variation of the magnetic contrast. It is stronger along the center of the element than at the edges. In particular, in the corners the domain walls are more strongly pinned and the rotation angle is smaller. This example with the competition of incoherent and coherent rotation processes shows clearly that the magnetodynamics is governed by the torque-induced processes rather than energy (stray field) minimization principles.

Instead of these incoherent rotation events, one may also observe a quite different behavior, if the same type of samples is subjected to much shorter (subnanosecond) magnetic field pulses. Choe *et al.* noted a gyrotropic rotation of the vortex cores with frequencies in the MHz range [67], whereas Quitmann *et al.* found a linear motion of the vortex, similar to the situation in Fig. 21 [68]. The reason for this difference may be sought in the different pulse repetition frequencies (125 MHz [67] to 62.5 MHz [68]) of both experiments, leading to different states of relaxation between the pulses. This is only a glimpse of the wealth of dynamic phenomena accessible by time-resolved EEM.

6 Application to Functional Materials II: Nonmagnetic Systems

6.1 Redox Processes in Resistive Oxides

The memristive behavior observed in many oxidic materials is generally related to electrically induced redox processes [69, 70]. The microscopic mechanisms, however, are still under discussion and subject to current investigations. Of particular importance are two aspects: *(i)* the atomic nature of the chemical changes related to the redox process, and *(ii)* the size of the region where these changes occur. Depending on the material in question, the proposed mechanisms range from homogeneous conductivity changes in the bulk down to a very local formation of conductive filaments with nanometer diameter.

One of the materials where resistive switching is observed is highly non-stoichiometric amorphous Gallium Oxide (a-GaOx) [71]. The microscopic mechanism is related to a metal-insulator transition involving a local formation of crystalline Ga_2O_3 within the metastable oxide matrix. The results of a microspectroscopic investigation with XPEEM are compiled in Fig. 22 [72].

The experiment used soft x-rays and due to the limited information depth (cf. 5.1) a particular preparation procedure was involved, which is sketched in Fig. 22a. In a first step, contact areas on the bare *a*-GaOx surface were defined by Pt top electrodes (TE). The common bottom electrode was provided by ITO. By applying a voltage to a contact, the material underneath was polarized into the high resistive (HRS) or low resistive state (LRS). For some contacts several sweep cycles were performed to study forming processes. This preswitching procedure resulted in a set of contacts in different switching states. After the polarization, in a second step UV set epoxy resin was pasted onto the top electrodes. Subsequently the top electrodes were removed by scratching off the epoxy resin thereby exposing the GaOx surface. Fig. 22b shows a an energy-filtered PEEM image in the Ga *3d* binding energy region, measured at the border of the pristine surface (bright grey region with red square) and the bare surface of the HR film area which has been subjected to a +2 V-bias during the preswitching step (diameter of the original contact 200 µm) (dark grey region with blue square). The black line appearing along the border is due to segregation of contaminants such as Pt and resin residues. The inset in the figure displays local XPS spectra which have been reconstructed from the PEEM images at the red point within the pristine surface area (red line) and at the blue point within the polarized surface area (blue line).

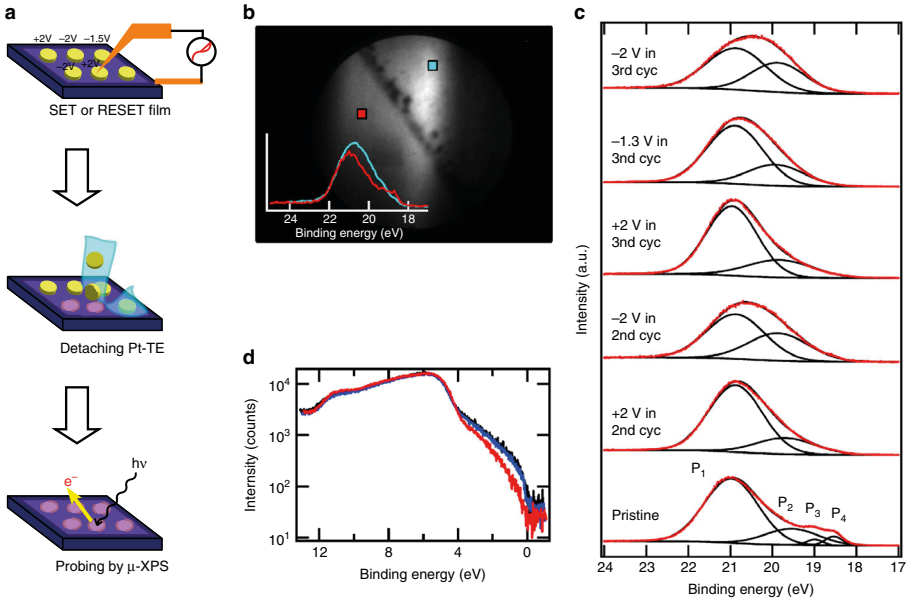


Fig. 22: *Ga electronic state change through resistive switching of a-GaOx device. (a) Preparation procedure of the bare a-GaOx surface after polarization to HRS or LRS. (b) Energy-filtered PEEM image in the Ga 3d binding energy region. Inset: Reconstructed XPS spectra. (c) Micro-XPS in the Ga 3d binding energy region measured on the pristine surface and at regions of the top electrode TE polarized at different voltages. (d) Near-Fermi edge spectra of the a-GaOx film: pristine (black), HRS (red) and LRS (blue). See text for further details. From [72].*

The micro-XPS spectra form the basis of an analysis of the chemical changes due to the resistive

switching. Fig. 22c compares spectra in the Ga $3d$ binding energy region measured on the pristine surface and at locations of the removed top electrodes, which were polarized at different voltages during the 2nd and 3rd sweep cycle. In the polarized films, the measured spectrum (red dots) can be described well by two peaks (black line) corresponding to Ga^{3+} (P1) and Ga^{+} (P2). From this result we find that Ga is highly reduced. A couple of peaks corresponding to metallic Ga^{0} (P3 and P4) are found only in the unpolarized pristine film. The molar ratios of Ga^{3+} /Ga^{+} are 0.16/0.84 at 2 V and 0.52/0.48 at -2 V in the 2nd sweep cycle and 0.19/0.81 at 2 V, 0.28/0.71 at -1.3 V and 0.52/0.48 at -2 V in the 2nd sweep cycle and 0.19/0.81 at 2 V, 0.28/0.71 at -1.3 V and 0.50/0.50 at -2 V in the 3rd sweep cycle, as calculated by a peak fitting analysis.

The change between the HRS and LRS state should also show up in the valence electronic structure. Indeed, valence band spectra close to the Fermi edge E_F reveal a distinct change in the density of states (Fig. 22d). The spectral weight at the Fermi edge in the HRS state (at $+2$ V, red) is markedly reduced, as compared to the specta from the pristine *a*-GaOx film (black) and LRS (at -2 V, blue). This indicates a reduction of the states at E_F available for electrical transport. These results demonstrate that the density of states near the top of the valence band is variable by negative or positive biasing, thereby changing the resistivity in the contact area. Further investigations reveal that the resistivity changes involve both oxygen vacancies and electrons and take place homogeneously over the entire contact area [72].

6.2 Overcoming the Information Depth Barrier

The above described approach of top electrode removal has a serious shortcoming, as it does only allow the study of preswitched systems. This precludes *in-operando* experiments during the resistive switching process or time-resolved investigations, which need the top electrode to be in place. Considering a minimum thickness of a working electrode of about 5 - 10 nm, excitation with soft x-rays will not permit one to probe the material underneath the top electrode. In order to overcome the limits imposed by the photoelectron attenuation length in the solid, we have to increase the kinetic energy of the photoelectrons into the keV regime. For kinetic energies of 10 keV the inelastic mean free path of photoelectrons in metals may reach up to 10 nm [73]. This requires an excitation energy in the hard x-ray regime. The main question, whether energy-filtered PEEM can yield reasonable data under these conditions has been successfully answered in 2012 [74]. The instrument used for this purpose was a modified electrostatic PEEM with a double-hemispherical electron analyzer. The acceleration voltage of the objective lens was increased to 30 kV in order to preserve the imaging properties also at high kinetic electron energies. In order to compensate for the low transmission of the electron optics at high kinetic energies and the low photoexcitation cross sections for hard x-ray excitation the microscope was operated at the largest contrast aperture (diameter 500 µm) thereby limiting the lateral resolution to a few hundred nm.

Results of such measurements for a test pattern are shown in Fig. 23. The test pattern consists of an arrangement of Au squares lithographically defined on a Si wafer. The left hand side of the figure reproduces a survey photoemission spectrum taken with $h\nu = 4900$ eV photon energy. The Au valence band (VB) and various Au and Si core levels are clearly visible. The region of the Si $2p$ and Au $4f$ levels has ben mapped at $h\nu = 6500$ eV photon energy in more detail, revealing the spin-orbit split Au levels. These lines (Si $2p$ and Au $4f_{7/2}$) have been employed

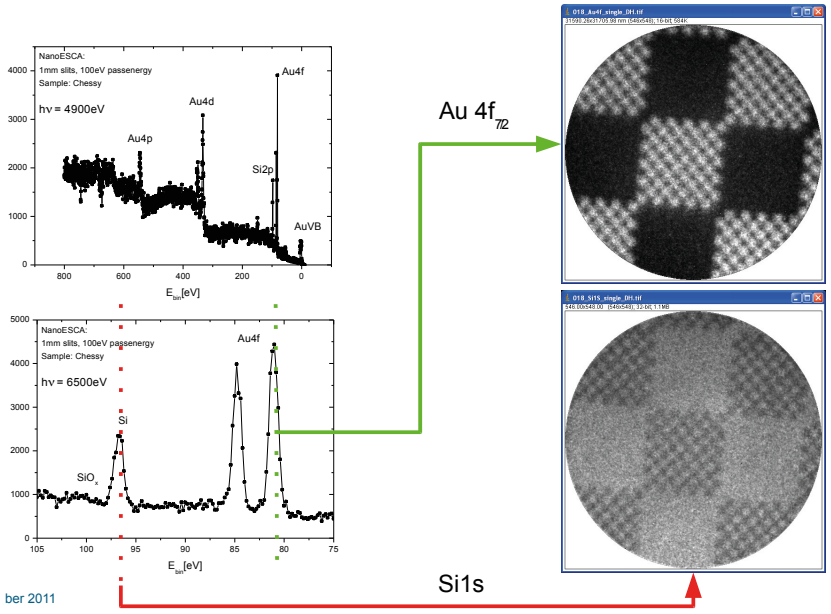


Fig. 23: *Hard x-ray energy-filtered PEEM on a Au/Si test structure. Left: XPS spectra taken at photon energies of* $h\nu = 4900$ *eV and* $h\nu = 6500$ *eV, showing the Si* $2p$ *and Au* $4f$ *core levels. Right: PEEM images recorded at the respective core levels from Si and Au. The pattern consists of smaller squares of* $1 \times 1\mu\mathrm{m}^2$ *which are grouped into larger squares of* $10 \times 10\mu\mathrm{m}^2$.

to acquire energy-filtered images of the surface, shown on the right hand side of Fig. 23. As expected the image contrast inverts between the two core levels. The smallest squares visible have a size of $1 \times 1\mu\mathrm{m}^2$. The lateral resolution obtained in this experiment was about 400 nm, which is compatible with the size of the contrast aperture chosen.

The issue of information depth is addressed by means of another specially designed test sample (Fig. 24a). For this purpose a trench in a Si wafer has been filled by Au. The resulting system has been chemomechanically polished, resulting in a rather sharp Si/Au boundary and a flat surface, onto which a Cr film with a thickness gradient ("wedge") has been deposited. Again the experiment imaged the lateral distribution of the $4f_{7/2}$ photoelectrons at $h\nu = 6500$ eV photon energy. The three-dimensional plot of the Au photoemission signal shows the gradual attenuation through the Cr wedge in one direction and the sharp cut-off due to the Au/Si boundary in the orthogonal direction (Fig. 24b). From these data we can easily determine the electron attenuation length λ (EAL) [75]. For the Au $4f_{7/2}$ photoelectrons which have a kinetic energy of $E_{kin} \simeq 6460$ eV we find $\lambda \simeq 8$ nm (Fig. 24c). Similar experiments at the Au $3d_{5/2}$ ($E_{kin} \simeq 4340$ eV) yield $\lambda \simeq 5$ nm.

These results demonstrate that PEEM with hard x-ray excitation has the potential for *in-operando*

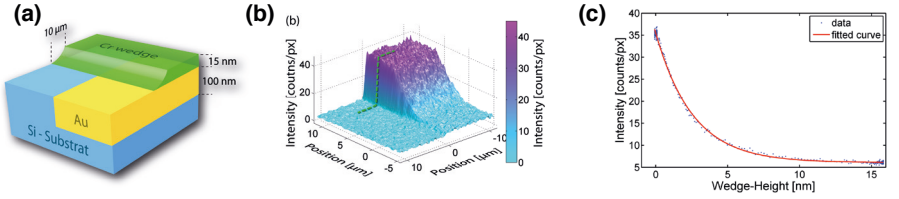


Fig. 24: *HAXPEEM on a* $\{\mathrm{Si} \mid \mathrm{Au}\}/\mathrm{Cr}$*-wedge test sample. (a) Structure of the sample. (b) Lateral distribution of the Au* $4f_{7/2}$ *photoelectrons mapped at* $h\nu = 6500$ *eV photon energy. (c) Electron attenuation length extracted from the damping of the Au* $4f_{7/2}$ *signal in the Cr overlayer. The fit is a simple exponential function. From [75].*

studies of resistive switching systems. A necessary requirement, however, is the availability of high-brilliance hard x-ray radiation from a dedicated synchrotron radiation source.

6.3 Probing the Photoelectron Spin

Novel functional materials and concepts for spintronics rely on spin-orbit coupling rather than exchange interaction. Topological insulators are nonmagnetic materials which exhibit a peculiar band inversion. The band inversion creates topologically protected spin-polarized surface states, the spin texture of which depends strongly on the wave vector $\boldsymbol{k}$ [76]. Examples for topological insulators are Bi_2Se_3, Bi_2Te_3 or Sb_2Te_3. The polarized surface states are considered as potential sources for spin-polarized currents. For further details on spin-orbit related spin-polarized electronic states see the contribution **C5** by Plucinski.

The experimental challenge with topological insulators is a mapping of the spin texture in the electronic states. This is normally achieved in a spin-resolved ARPES set-up (c.f. **C5**), but this approach is not very efficient, mainly for two reasons. First, the spin detectors currently in use are typically single channel detectors, i.e. only a single spectrum can be measured at a time. The two-dimensional mapping capability of modern display analyzers cannot be used in this mode. Second, the spin detection mechanism itself involves a scattering process which reduces the signal by 3 - 4 orders of magnitude [77]. As a consequence, detailed spin-resolved ARPES experiments have been very time-consuming in the past. The situation has been improved only very recently due to the evolution of the PEEM optics in combination with the development of highly efficient two-dimensional spin detectors.

Considering the properties of the PEEM optics it is important to realize that the objective lens can be operated in two different imaging modes: (i) real-space imaging, and (ii) reciprocal space imaging. The first mode has been extensively described and used above. In the second mode, the angular distribution of the emitted photoelectrons is magnified and projected onto the image detector. By means of the energy filter a cut through the Brillouin zone at a defined kinetic energy can be taken and one obtains a two-dimensional photoelectron intensity distribution $I(E_{kin}, k_x, k_y)$ in a single measurement [78]. By scanning the kinetic energy a map of the entire Brillouin zone can be recorded slice-by-slice. Depending on the settings of the objective

lens and the photon energy, also states beyond the first Brillouin zone can be probed. This special mode of operation is now known as *k-space microscopy* or *momentum microscopy* and is increasingly used to map electronic structures.

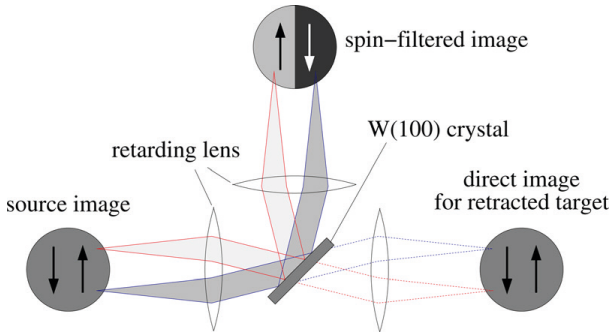


Fig. 25: *Principle of a two-dimensional spin polarization detector involving spin-dependent scattering on a W(100). The source image from a schematic magnetic domain pattern is projected onto the scattering target and transfered to a second image detector at an angle of 90°. For spin-integrated measurements the scattering crystal is retracted and the image is transfered to the straight-through image detector. From [79].*

As we have already mentioned above, the physical mechanism behind the spin polarization analysis involves a scattering process of the electrons. This scattering process "converts" the spin polarization along a quantization axis into an intensity signal. The quantization axis is defined by the scattering geometry [36]. The conversion efficiency, the spin-polarization sensitivity S, depends sensitively on the scattering angle and the scattering energy, both of which have to be carefully controlled. Behind the energy filter of a PEEM the electrons forming the image are quasi monoenergetic. This property forms the basis of a two-dimension spin detector concept, which is sketched in Fig. 25 [79]. The monoenergetic electron beam leaving the energy filter is guided onto a scattering target – in the present case a clean W(100) single-crystal surface – and scattered at an angle of 45°. The strong spin-orbit coupling in W leads to a high spin-selectivity of the scattering process for a specific kinetic energy of $E_{scat} = 27$ eV. The spin quantization axis is oriented perpendicular to the scattering plane and the spin sensitivity is about 20%. Therefore, one observes an intensity modulation in the image of the scattered electrons under 90°, which is directly related to the spin polarization in the original image. The scheme in Fig. 25 refers to the real-space imaging of magnetic domain structure with up and down domains. As the photoelectrons from a ferromagnetic domain are highly spin-polarized, the spin-dependent scattering at the W(100) target leads to in a bright/dark contrast in the spin-resolved image. For spin-integrated measurements the scattering target is simply retracted allowing the electrons to directly reach the image detector in the straight-through direction. The information on the magnetic domain structure that can be gained in this way is similar to that obtained in an MXCD experiment (cf. Chapter 5.1).

The major application of the two-dimensional spin-detector concept, however, comes with the momentum microscopy. This is illustrated in Fig. 26 showing spin-resolved photoemission re-

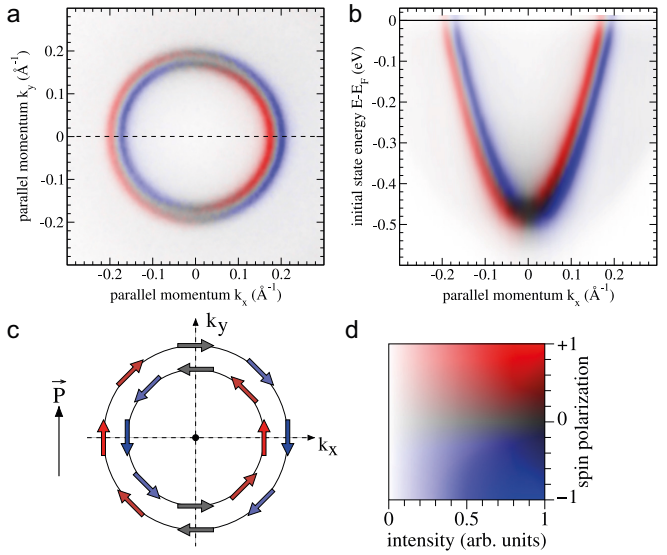


Fig. 26: *Photoemission from the Au(111) surface state with p-polarized 6.05 eV photons. (a) Measured spin polarization and intensity map at E_F (2D color code see (d)). (b) Spin resolved dispersion along the horizontal ($k_y = 0$) axis. (c) Schematic model of the spin texture of the Rashba surface state. Arrows indicate the spin direction, the color corresponds to the observed projection on the quantization axis $\vec{P}$. From [80].*

sults from the Shockley surface state on the Au(111). This surface state is known to exhibit a distinct splitting due to Rashba interaction (for details, cf. **C5**). The parabolic dispersion around the Fermi energy E_F is well-reproduced in the data in Figs. 26a and b. These plots combine intensity (shading) and spin polarization information (color), whereby blue and red correspond to spin-down and spin-up, respectively. In the $\{E, k_x, k_y = 0\}$ cut we find two parabolas with distinct spin-up and spin-down character. At the center of the Brillouin zone the spin polarization is reduced to hybridization of both parabolas. In the $\{E = 0, k_x, k_y\}$ the surface state dispersion forms two concentric circles with a continuous, but anticyclical variation of the spin polarization signal along the circle. This clearly proves that the spin polarization vector changes from k-point to k-point. The experimentally determined spin texture of the surface state is compatible with the schematic model depicted in Figs. 26c. These experiments were performed at an energy resolution of 12 meV, measured at the Fermi edge of the helium-cooled Au(111) sample. The instrumental momentum resolution of 0.005 Å^{-1}is among the best values for state-of-the-art spin-integrated photoemission experiments.

The example above demonstrates the potential and the advantages of spin-resolved momentum microscopy over a conventional spin-resolved photoemission experiment with single-channel spin-detection.

7 Concluding Remarks

These lecture notes can give only a brief introduction into the XPEEM technique and cover only a limited selection of physical phenomena which can be investigated with this approach. For a more concise information on PEEM the reader is referred to a number of recent review articles [14, 19, 30, 63] and the book of E. Bauer [11]. In future studies, two major issue will be of importance: *(i)* improvement of the lateral resolution by corrected electron optics; *(ii)* further improvement of the information depth; and *(iii)* in-operando and time-resolved studies. The investigations associated with the last issue will make use of the intrinsic time structure of the synchrotron radiation generated in storage rings. The ultimate goal will be the laterally resolved study of resistive switching processes on a nanosecond or even picosecond time scale.

References

[1] *Chips 2020 – A Guide to the Future of Nanoelectronics*, eds. B. Hoefflinger (Springer, Berlin, 2012).

[2] *Green IT: Technologies and Applications*, eds. J. H. Kim and M. J. Lee (Springer, Berlin, 2011).

[3] see, for example, the website *http://www.hitachigst.com/hdd/research/*.

[4] L. Reimer: *Scanning Electron Microscopy* (Springer-Verlag, Berlin, 1985).

[5] T. Warwick, K. Franck, J.B. Kortwright, G. Meigs, M. Moronne, S. Myneni, E. Rotenberg, S. Seal, W. F. Steele, H. Ade, A. Garcia, S. Cerasari, J. Denlinger, S. Hayakawa, A.P. Hitchcock, T. Tyliszczak, E.G. Rightor, H.-J. Shin and B. Tonner, Rev. Sci. Inst. **69** (1998) 2964.

[6] *Scanning Tunneling Microscopy*, eds. H.-J. Güntherodt and R. Wiesendanger (Springer-Verlag, Berlin, 1992).

[7] K. Koike and K. Hayakawa, Jpn. J. Appl. Phys. **23** (1984) L187.

[8] Y. Martin and H.K. Wickramasinghe, Phys. Rev. Lett. **50** (1987) 1455.

[9] M. Bode, M. Getzlaff and R. Wiesendanger, Phys. Rev. Lett. **81** (1998) 4256.

[10] E. Bauer and W. Telieps, in: *Emission and low energy reflection electron microscopy*, eds. A. Howie and U. Valdré (Plenum Press, New York, 1988).

[11] E. Bauer, *Surface Microscopy with Low Energy Electrons* (Springer, New York, 2014).

[12] M. Brüche, Z. f. Naturforschung, **11** (1934) 287.

[13] L.H. Veneklasen, Rev. Sci. Instrum. **63** (1992) 5513.

[14] G. Schönhense, J. Phys.: Condens. Matter **11** (1999) 9517.

[15] C.M. Schneider and G. Schönhense, Rep. Prog. Phys. **65** (2002) 1785.

[16] R. Fink, M.R. Weiss, E. Umbach, D. Preikszas, H. Rose, R. Spehr, P. Hartel, W. Engel, R. Degenhardt, R. Wichtendahl, H. Kuhlenbeck, W. Erlebach, K. Ihmann, R. Schlögl, H.-J. Freund, A.M. Bradshaw, G. Lilienkamp, T. Schmidt, E. Bauer and G. Benner, J. Electron Spectrosc. Relat. Phenom. **84** (1997) 231.

[17] G.F. Rempfer, W.P. Skoczylas and O.H. Griffith, Ultramicroscopy **36** (1991) 196.

[18] S. Anders, H.A. Padmore, R.M. Duarte, T. Renner, T. Stammler, A. Scholl, M.R. Scheinfein, J. Stöhr, L. Séve and B. Sinkovic, Rev. Sci. Instrum. **70** (1999) 3973.

[19] S. A. Nepijko, N.N. Sedov and G. Schönhense, Adv. Imag. Electron Phys. **113** (2000) 205.

[20] see: `http://www.elmitec.de`, respective instrument *PEEM III with analyzer*

[21] see: `http://www.specs.de`, respective instrument *FE-LEEM P90*

[22] see: `http://www.scientaomicron.com/en/products/nanoesca-/instrument-concept`.

[23] M. Escher, N. Weber, M. Merkel, C. Ziethen, P. Bernhard, G. Schönhense, S. Schmidt, F. Forster, F. Reinert, B. Krömker and D. Funnemann, J. Phys.: Condens. Matt. **17** (2005) S1329.

[24] R. Wichtendahl, R. Fink, H. Kuhlenbeck, D. Preikszas, H. Rose, R. Spehr, P. Hartel, W. Engel, R. Schlögl, H.-J. Freund, A.M. Bradshaw, G. Lilienkamp, Th. Schmidt, E. Bauer, G. Benner, E. Umbach, Surf. Rev. Lett. **5** (1998) 1249.

[25] A. Kiejna and K.F. Wojciechowski, Prog. Surf. Sci. **11** (1981) 293.

[26] O. Renault, R. Brochier, P.-H. Haumesser, N. Barrett, B. Kromker, and D. Funnemann, e-J. Surf. Sci. Nanotech. **4** (2006) 431.

[27] T. Schmidt, S. Heun, J. Slezak, J. Diaz, K.C. Prince, G. Lilienkamp, and E. Bauer, Surf. Rev. Lett. **5** (1998) 1287.

[28] W. Swiech, G.H. Fecher, C. Ziethen, O. Schmidt, G. Schönhense, K. Grzelakowski, C.M. Schneider, R. Frömter and J. Kirschner, J. Electron Spectr. Rel. Phen. **84** (1997) 171.

[29] C. Ziethen, O. Schmidt, G.K.L. Marx, G. Schönhense, R. Frömter, J. Gilles, J. Kirschner, C.M. Schneider, and O. Gröning, J. Electron Spectr. Rel. Phen. **107** (2000) 261.

[30] J. Stöhr and S. Anders, IBM J. Res. Develop. **44** (2000) 535.

[31] G.V. Spivak, T.N. Dombrovskaia and N.N. Sedov, Sov. Phys. Dokl. **2** (1957) 120.

[32] G. v. d. Laan, B. T. Thole, G. A. Sawatzky, J. B. Goedkoop, J. C. Fuggle, J. M. Esteva, R. Karnatak, J. P. Remeika and H. A. Dabkowska, Phys. Rev. B **34** (1986) 6529.

[33] G. Schütz, W. Wagner, W. Wilhelm, P. Kienle, R. Zeller, R. Frahm and G. Materlik, Phys. Rev. Lett. **58** (1987) 737.

[34] C.T. Chen, F. Sette, Y. Ma and S. Modesti, Phys. Rev. B **42** (1990) 7262.

[35] T. Nakagawa, T. Yokoyama, M. Hosaka and M. Katoh, Rev. Sci. Instrum. bf 78 (2007) 023907.

[36] J. Kessler, *Polarized Electrons* (Springer-Verlag, Berlin, 1985).

[37] C.M. Schneider, K. Holldack, M. Kinzler, M. Grunze, H.P. Oepen, F. Schäfers, H. Petersen, K. Meinel and J. Kirschner, Appl. Phys. Lett. **63** (1993) 2432.

[38] J. Stöhr, Y. Wu, M.G. Samant, B.D. Hermsmeier, G. Harp, S. Koranda, D. Dunham and B.P. Tonner, Science **259** (1993) 658.

[39] R. Nakajima, J. Stöhr and Y. U. Idzerda, Phys. Rev. B **59** (1999) 6421.

[40] A. Hubert and R. Schäfer, *Magnetic Domains* (Springer-Verlag, Berlin, 1998).

[41] C.M. Schneider, R. Frömter, C. Ziethen, W. Swiech, N.B. Brookes, G. Schönhense and J. Kirschner, Mat. Res. Soc. Symp. Proc. **475** (1997) 381.

[42] H.P. Oepen and J. Kirschner, Phys. Rev. Lett. **62** (1989) 819.

[43] M.R. Scheinfein, J. Unguris, R.J. Celotta and D.T. Pierce, Phys. Rev. Lett. **63** (1989) 668.

[44] S. Chikazumi, *Physics of Magnetism* (Wiley & Sons, New York, 1964).

[45] R.W. DeBlois and J.C.D. Graham, J. Appl. Phys. **29** (1958) 931.

[46] H.P. Oepen and J. Kirschner, Scanning Microscopy **5** (1991) 1.

[47] J. Hunter Dunn, D. Arvanitis, N. Mårtensson, M. Tischer, F. May, M. Russo and K. Baberschke, J. Phys.: Cond. Matt. **7** (1995) 1111.

[48] J. Unguris, R.J. Celotta and D.T. Pierce, Phys. Rev. Lett. **67** (1991) 140.

[49] W. Kuch, R. Frömter, J. Gilles, D. Hartmann, C. Ziethen, C.M. Schneider, G. Schönhense, W. Swiech, and J. Kirschner, Surf. Rev. Lett. **5** (1998) 1241.

[50] C.M. Schneider, K. Meinel, J. Kirschner, M. Neuber, C. Wilde, M. Grunze, K. Holldack, Z. Celinski, and F. Baudelet, J. Magn. Magn. Mater. **162** (1996) 7.

[51] D. Alders, L. H. Tjeng, F. C. Voogt, T. Hibma, G.A. Sawatzky, C.T. Chen, J. Vogel, M. Sacchi and S. Iacobucci, Phys. Rev. B **57** (1998) 11623.

[52] D. Spanke, V. Solinus, D. Knabben, F.U. Hillebrecht, F. Ciccacci, L. Gregoratti and M. Marsi, Phys. Rev. B **58** (1998) 5201.

[53] J. Stöhr, A. Scholl, T.J. Regan, S. Anders, J. Lüning, M.R. Scheinfein, H.A. Padmore and R.L. White, Phys. Rev. Lett. **83** (1999) 1862.

[54] F. Nolting, A. Scholl, J. Stöhr, J. Fompeyrine, H. Siegwart, J.-P. Locquet, S. Anders, J. Lüning, E.E. Fullerton, M.F. Toney, M.R. Scheinfein and H.A. Padmore, Nature **405** (2000) 767.

[55] F.U. Hillebrecht, H. Ohldag, N.B. Weber, C. Bethke, U. Mick, M. Weiss, and J. Bahrdt, Phys. Rev. Lett. **86** (2001) 3419.

[56] D. Alders, J. Vogel, C. Levelut, S.D. Peacor, T. Hibma, M. Sacchi, L.H. Tjeng, C. T. Chen, G. van der Laan, B.T. Thole and G.A. Sawatzky, Europhys. Lett. **32** (1995) 259.

[57] H. Komatsu and M. Ishigame, J. Mat. Sci. **20** (1985) 4027.

[58] J. Nogues and I.K. Schuller, J. Magn. Magn. Mater. **192** (1999) 203.

[59] T.C. Schulthess and W.H. Butler, Phys. Rev. Lett. **81** (1998) 4516.

[60] *Handbook on Synchrotron Radiation*, eds. D. Eastman and Y. Farge (North-Holland, Amsterdam, 1983).

[61] A. Krasyuk, A. Oelsner, S. Nepijko, A. Kuksov, C.M. Schneider and G. Schönhense, Appl. Phys. A **76** (2003) 836.

[62] J. Vogel, W. Kuch, M. Bonfim, J. Camarero, Y. Pennec, F. Offi, K. Fukumoto, J. Kirschner, A. Fontaine and S. Pizzini, Appl. Phys. Lett. **82** (2003) 2299.

[63] G. Schönhense, H.-J. Elmers, S.A. Nepijko, and C.M. Schneider, Adv. Imag. Electron Phys. **142** (2006) 160.

[64] W.K. Hiebert, G.E. Ballentine, L. Lagae, R.W. Hunt and M.R. Freeman, J. Appl. Phys. **92** (2002) 392.

[65] C.M. Schneider, A. Kuksov, A. Krasyuk, A. Oelsner, D. Neeb, S.A. Nepijko, G. Schönhense, J. Morais, I. Mönch, R. Kaltofen, C. de Nadaï and N.B. Brookes, Appl. Phys. Lett. **85** (2004) 2562.

[66] D. Neeb, A. Krasyuk, A. Oelsner, S.A. Nepijko, H.J. Elmers, A. Kuksov, C.M. Schneider and G. Schönhense, J. Phys.: Condens. Matt. **17** (2005) S1381.

[67] S.-B. Choe, Y. Acremann, A. Scholl, A. Bauer, A. Doran, J. Stöhr and H.A. Padmore, Science **304** (2004) 420.

[68] J. Raabe, C. Quitmann, C.H. Back, F. Nolting, S. Johnson and C. Buehler, Phys. Rev. Lett. **94** (2005) 217204.

[69] R. Waser and M. Aono, Nat Mater **6** (2007) 833.

[70] R. Waser and M. Wuttig, Adv. Funct. Mater. **25** (2015) 6285.

[71] L. Nagarajan, R. A. De Souza, D. Samuelis, I. Valov, A. Borger, J. Janek, K.-D. Becker, P. C. Schmidt and M. Martin, Nat Mater **7** (2008) 391.

[72] Y. Aoki, C. Wiemann, V. Feyer, H.-S. Kim, C. M. Schneider, H. Ill-Yoo and M. Martin, Nat Commun **5** (2014) 3473.

[73] S. Tanuma, C. J. Powell and D. R. Penn, Surf. Interf. Anal. **43** (2011) 689.

[74] C. Wiemann, M. Patt, S. Cramm, M. Escher, M. Merkel, A. Gloskovskii, S. Thiess, W. Drube and C. M. Schneider, Appl. Phys. Lett. **100** (2012) 223106.

[75] M. Patt, C. Wiemann, N. Weber, M. Escher, A. Gloskovskii, W. Drube, M. Merkel and C. M. Schneider, Rev. Sci. Instrum. **85** (2014) 113704.

[76] M. Z. Hasan and C. L. Kane, Rev. Mod. Phys. **82** (2010) 3045.

[77] J. Kirschner, in: *Sources and Detectors for Spin Polarized Electrons*, eds. R. Feder (World Scientific, Singapore, 1985)

[78] A. Winkelmann, C. Tusche, A. A. Ünal, M. Ellguth, J. Henk and J. Kirschner, New J. Phys. **14** (2012) 18.

[79] C. Tusche, M. Ellguth, A. A. Ünal, C. Chiang, A. Winkelmann, A. Krasyuk, M. Hahn, G. Schönhense and J. Kirschner, Appl. Phys. Lett. **99** (2011) 032505.

[80] C. Tusche, A. Krasyuk and J. Kirschner, Ultramicroscopy **159** (2015) 520.

C 7 Scanning Probe Microscopy

Ph. Ebert, M. Moors
PGI-5, PGI-7
Forschungszentrum Jülich

Contents

1 Introduction **2**

2 The scanning tunneling microscope **3**
2.1 Theoretical fundamentals 3
2.2 Operating modes 8
2.3 Experimental realization 9
2.4 Applications of the classical scanning tunneling microscope 11

3 The scanning force microscope **17**
3.1 Theoretical principles of the scanning force microscope 18
3.2 Operation principle of scanning force microscope 19
3.3 Technical realization of a scanning force microscope 21
3.4 Applications of the scanning force microscope 22

4 Application of scanning probe microscopy in memristive cells **25**
4.1 Local conductivity atomic force microscopy (LC-AFM) 26
4.2 Scanning tunneling microscopy 29

5 Summary **33**

1 Introduction

Since the invention of the scanning tunneling microscope by Binnig and Rohrer in 1982 [1], a variety of scanning probe microscopy (SPM) techniques have rapidly developed into increasingly important tools for surface physics and for the characterisation of surface structures. The surge in applications of scanning probe microscopes is primarily due to their unique ability to provide real space images with atomic resolution of surface structures. Furthermore, the different types of scanning probe microscopes provide the possibility of investigating electrical, topographic, optical, magnetic, and many other types of surface properties. Depending on the mode of operation on which the scanning probe microscopes are based, they can be used for conducting and non-conducting as well as hard and soft materials.
A scanning probe microscope works according to a simple principle (Fig. 1): A probe is scanned over the surface of interest at a small distance, where an interaction between the probe and the surface is present. This interaction can be of various nature (electrical, magnetical, mechanical, etc.) and provides the measured signal (tunnel current, force, etc). Depending on the quality and type of probe, the measured signal can be observed to reproducibly vary at atomic distances during scanning of the surface. If these scanning processes are put together line by line, an ”image” of the surface is obtained. Such an image shows the spatial variation of the measured parameter, e.g., of the tunnel current.

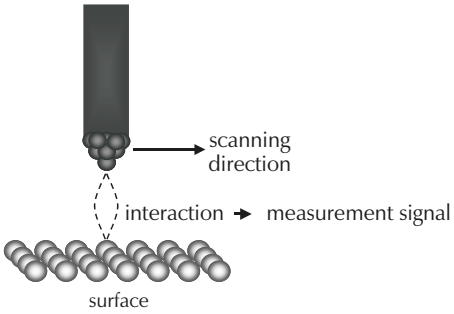


Fig. 1: *Principle of a scanning probe microscope. A probe tip is scanned over a surface. The interaction with the surface yields a signal which is used to derive a real space image of the surface.*

Of all the scanning probe microscopes the first invented scanning tunneling microscope (STM) still provides the highest routinely achieved resolution. It is used for semiconductors, metals, and superconductors, because it constitutes a probe for electrical properties and thus requires electrically conducting surfaces. It measures the tunnel current between a metallic, extremely fine tip and the surface. Thereby the STM probes the local density of surface states, whose atomic-scale variations allows to image surfaces with atomic resolution. Fig. 2 shows as example a STM image of the InP(110) surface. Each bright local peak represents the increased local density of states near an atom on the surface. Black holes indicate missing atoms, i.e., vacancies [2]. Thus, one can recognize atomic rows and individual point defects.
In addition to the tunnel current one can probe also the spin of the tunneling electrons using specially prepared magnetic tips. Thereby a spin-polarized scanning tunneling microscope (SP-STM) is obtained, which allows to probe the spin structure of surfaces [3].

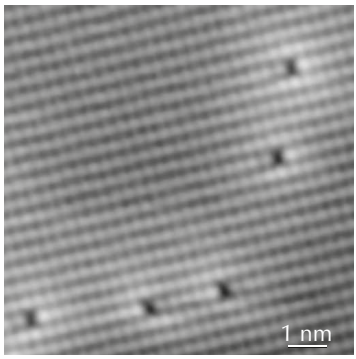


Fig. 2: *Scanning tunneling microscope image of several phosphorus vacancies on a n-doped InP(110) surface. The image was obtained at negative voltages applied to the sample. Thus, the electrons tunneled from the filled InP(110) surface states into empty tip states. Adapted with permission from [2], ©1994 American Physical Society.*

Another frequently used scanning probe microscope is the scanning force microscope (SFM, also known as atomic force microscope, AFM), which probes the force between a tip mounted on a special spring and the surface [4]. The force involved can arise from van der Waals, electrostatic, magnetic, or repulsive atomic interactions.
Since the initial development of these two basic scanning probe microscopes, a number of additional ones were invented, which basically only differ by probing different sample properties, such as temperature distributions (scanning thermal microscope), acoustic properties (scanning acoustical microscope), etc.. Of all those maybe the scanning near-field optical microscope (SNOM) [6] gained a larger importance, because it utilizes the special focusing of the optical near field at a pointed light guide to probe optical surface properties.
In this lecture, I will focus primarily of the two most representative scanning probe microscopes, the scanning tunneling and the scanning force microscopes and some of its particularly interesting variants. First, the principles of scanning probe microscopes are introduced using the scanning tunneling microscope as example. Each microscopes ability and potential is illustrated with selected examples of applications.

2 The scanning tunneling microscope

2.1 Theoretical fundamentals

A scanning tunneling microscope uses a fine metallic tip (called tunneling tip) as probe. A bias voltage is applied between the tip and the electrically conducting sample surface (see Fig. 3). Then the tip is approached toward the surface until a electrical current flows. This happens at tip-surface separations in the order of 0.5 to 1 nm. The current is a tunnel current based on the quantum-mechanical *tunnel* effect [7, 8, 9]. After a tunneling contact is established, the tip is moved laterally over the surface by a piezoelectric *scanning* unit, whose mechanical extension can be controlled by applying appropriate voltages. The scanning unit is typically capable of scanning an area of a few nm^2 up to several μm^2. Thereby a *microscopic* image of the spatial variation of the tunnel current is acquired. Hence the name *scanning tunneling microscope*.

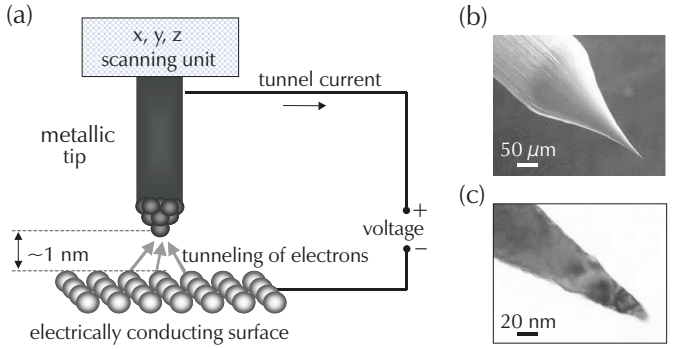


Fig. 3: *(a) Schematic drawing of a classical scanning tunneling microscope. The tunnel current is used as measuring signal. (b) and (c) show scanning and transmission electron microscope images, respectively, of a typical tungsten tips used for a classical scanning tunneling microscope with no spin sensitivity. Note the sharpness of the tips, which have a radius of curvature below 10 nm.*

At this stage we discuss what kind of atomic-scale structures can be made visible by utilizing the tunnel effect in the scanning tunneling microscope. These structures must by nature correspond to electrical states from or into which the electrons can tunnel. In the tunneling process, the electrons must tunnel through the vacuum gap between the tip and the sample surface. This vacuum gap represents a potential barrier, i.e. an energetically forbidden region. The tunnel effect allows a particle (here an electron) to tunnel through this potential barrier even though the electron's energy is lower than the barrier height. The probability of such a process decreases exponentially with the geometrical distance between the tip and the sample (determining the width of the potential barrier) and with increasing barrier height. An experimental apparatus making use of the tunnel effect must therefore minimize the width of the potential barrier to the degree that electrons can tunneled through it. This is realized in the scanning tunneling microscope configuration by moving the tip very close (about 1 nm or less) to the surface. The electrons can then pass between the surface and the tip. The direction of the *macroscopic* tunnel current is fixed by applying a voltage between sample and tip, even if electrons tunnel in both directions, but with different probabilities due to the applied voltage. Note, the tunneling process of an electron through an energetically forbidden region is instantaneously and thus, the tunneling electron does not stay a measurable time span in the forbidden potential barrier region [10].

In order to explain and interpret the images of the surface states obtained in this way, efforts to develop a theory were made soon after the invention of the scanning tunneling microscope. One of the possible theoretical approaches is based on Bardeen's idea of applying a transfer Hamiltonian operator to the tunneling process [11]. This had the advantage of adequately describing the many-particle nature of the tunnel junction. In the model, a weak overlap of the wave functions of the surface states of the two electrodes (tunneling tip and sample surface) is assumed to allow a perturbation calculation. The resulting current between two planar electrodes is then given by

$$I \sim \int_{-\infty}^{\infty} |M(E)|^2 \cdot \rho_{\text{tip}}(E - eV) \cdot \rho_{\text{sample}}(E) \cdot [f(E - eV) - f(E)]dE \qquad (1)$$

with $f(E)$ being the Fermi function, M the tunneling matrix element, ρ_{sample} and ρ_{tip} the density of states of the sample and tip, respectively, and E the energy of the density of states.
On this basis, Tersoff and Hamann developed a simple theory of scanning tunneling microscopy [12, 13]. By assuming that the tunneling tip can be approximated by a metallic s-orbital with its center at the position ***r***, as shown schematically in Fig. 4, they obtained for the tunnel current in a STM-like configuration:

$$I \sim V \cdot \rho_{\text{tip}}(E_{\text{F}}) \cdot \rho_{\text{sample}}(\overrightarrow{R}_{\text{tip}}, E_{\text{F}}) \tag{2}$$

In addition, is was assumed that low voltages V (i.e., much smaller than the work function) are applied in order to linearly approximate the voltage dependence (see below for the high voltage extension of Eq. 2). $\rho_{\text{tip}}(E_{\text{F}})$ is the density of states of the tip and $\rho_{\text{sample}}(\overrightarrow{R}_{\text{tip}},E_{\text{F}})$ is that of the sample surface at the center $\overrightarrow{R}_{\text{tip}}$ of the tip orbital and at the Fermi energy E_{F}. Equation 2 shows that at low voltage *the scanning tunneling microscope thus images the electronic density of states at the sample surface near the Fermi energy*. However, this result also means that the scanning tunneling microscope images do not directly show the atoms, but rather the electronic states bound to the atoms. If we recall Fig. 2 the maxima are thus the filled states localized above the phosphorus atoms on the InP(110) surface and the dark holes are missing states arising from vacancies at the surface.

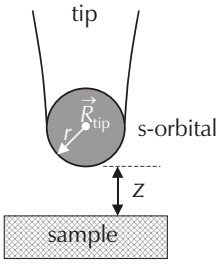


Fig. 4: *Schematic representation of the tunneling geometry used in the Tersoff-Hamann model. The tip approximated by a s orbital with a radius r at the position $\vec{R}_{tip}$.*

As can be seen in Eq. 2, the density of states of the probe tip enters in the measurement in the same way as the density of states of the sample. Thus depending on the exact density of states of the tip, the tunnel current will vary from tip to tip. It is therefore desirable to know the exact electronic state of the tip, but unfortunately, in actual experiments, every tip is different and the details remain almost always unknown, despite intense efforts to characterize the tips' apexes. Tip effect were nevertheless successfully distinguished from the real surface structure by careful measurements with a large number of tip configurations. Different tip configurations can be obtained during scanning over the surface by attracting individual atoms from the surface to the tip, as well as by special tip treatments, heating to high temperatures for cleaning in vacuum, ion sputtering, or field emission.
Equation 2 can be better interpreted by considering the exponential decay of the density of surface states into the vacuum with the effective inverse decay length κ_{eff}:

$$\kappa_{\text{eff}} = \sqrt{\frac{2m_{\text{e}}B}{\hbar^2} + \left|\boldsymbol{k}_{||}\right|^2} \tag{3}$$

m_e is the effective mass of the electron, $\boldsymbol{k}_{||}$ is the parallel wave vector of the tunneling electrons, which enters into the Eq. 3 due to the momentum conservation in the tunneling process [14]. B is the barrier height of the vacuum gap between the tip and the sample surface. The barrier height is a function of the applied voltage V and the work functions Φ_{sample} and Φ_{tip} of the sample and tip [15], respectively. It can be approximated to:

$$B = \frac{\Phi_{\text{tip}} + \Phi_{\text{sample}}}{2} - \frac{|eV|}{2} \tag{4}$$

The tunnel current thus decreases exponentially with the tip-sample distance z:

$$I \sim \exp\left[-2\kappa_{\text{eff}} z\right] \tag{5}$$

The exponential current–tip-sample distance dependence is essential for the high accuracy of a scanning tunneling microscope: First, very small changes in the tip-sample separation cause large changes in the tunnel current. This yields a high vertical resolution. Second, the tip just needs one nanotip, only about 0.1 nm closer to the surface than all other neighboring nanotips. Then essentially all the tunnel current flows only over this closest nanotip. Thus, even apparently wide and blunt tips can yield atomic resolution along the surface due to the exponential current-distance dependence and the presence of nanotips on the macroscopic tunneling tip.
The description of the tunnel current by Eq. 2, however, has an important restriction: it only applies to low voltages V, which multiplied by e must be much smaller that the work function ($eV \ll \Phi_{\text{sample}}$). This is reasonably correct for the tunneling conditions used to image metal surfaces. However, for the investigation of semiconductor surfaces, voltages of the order of 2 to 3 V, sometimes even larger as in case of GaN cleavage surfaces [16, 17] are required due to the existence of a wide band gap. Therefore the applied voltages times the electron charge e are in the same magnitude as the work function, and the above used approximations are insufficient. Thus the theory must be extended. The simplest extension yields:

$$I \sim \int_{E_{\text{F,tip}}}^{E_{\text{F,tip}}+eV} \rho_{\text{tip}}(W)\rho_{\text{sample}}(W + eV, \overrightarrow{R}_{\text{tip}})T(W, V)dW \tag{6}$$

$T(W,V)$ is a transmission coefficient, which depends on the energy of the electrons and the applied voltage. The transmission coefficient arises from the increased tunneling probability for surface states with smaller ionization energy (leading to a smaller effective tunneling barrier) and for the voltage dependence of the tunneling barrier. The transmission coefficient can be well approximated, if one considers the exponential decay of the density of states into the vacuum and the fact that the tunnel current is based of the density of states of the sample at the position of the tip $\overrightarrow{R}_{\text{tip}}$. This position corresponds to a tip-sample separation z and thus the transmission coefficient describes the z dependence of the density of states for a given energy and voltage. In case of positive voltages and zero parallel wave vector (tunneling from the Γ point) $T(W, V)$ can be thus approximated by [18]:

$$T(W, V) = \exp\left[-2z \cdot \sqrt{\frac{2m_e\left[\frac{\Phi_{\text{tip}}+\Phi_{\text{sample}}}{2} + \frac{eV}{2} - W\right]}{\hbar^2}}\right] \tag{7}$$

The tunnel current is thus composed of the product of the density of states of the tip and sample at all the different electron energies that are allowed to participate in the tunneling process (Fig. 5). For example, an image measured at -2 V applied to the sample, consequently shows all occupied sample states with an energy between the Fermi energy and 2 eV below the Fermi energy. In analogy tunneling at a positive voltages applied to the sample provides a measurement of the empty surface states in an energy interval determined again by the voltage.

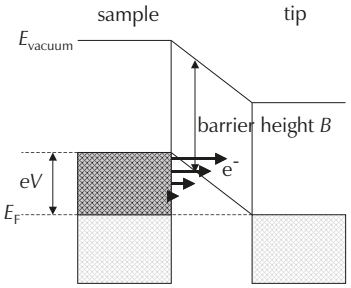


Fig. 5: *At high voltages not only the states near the Fermi energy E_F contribute to the current but all states whose energy ranges between E_F and E_F+eV.*

This effect can be illustrated further using the InP(110) surface, which has two surface states: an occupied state below the valence band edge and an empty state above the conduction band edge (Fig. 6). All other states are located geometrically deeper in the crystal or energetically deeper in the bands. They thus contribute only at high voltages [19], which will not be considered here. In the special case of the InP(110) surface, the occupied surface state is spatially located above the P atoms, whereas the empty state is bound to the In atoms (Fig. 6c1,c2). The P and In atoms are alternately arranged in zigzag rows. At negative sample voltages, the scanning tunneling microscope probes the occupied states located at the P sublattice, whose electrons tunnel into the empty states of the tunneling tip (Fig. 6a). Conversely, only the empty surface states at the In sublattice are probed at positive voltages applied to the sample (Fig. 6b) [19]-[21]. If the voltage polarity is changed every scan line, i.e. the occupied and the empty states are probed each alternating scan line, the two resulting images can be superimposed and the zigzag rows of alternating indium and phosphorus atoms become visible (Fig. 6c3).
Apart from the spatial distribution of the density of states, its energy dependence can be determined from current-voltage characteristics using Eq. 6. In order to do so, however, information is required about the transmission coefficient, which turns out to be a great obstacle even if approximations [22] are used. Therefore, in most cases, an experimentally viable approach is used, in which the density of states is approximated to [23] [24]:

$$\rho_{\text{sample}}(eV) \approx (dI/dV)/(\overline{I/V}) \tag{8}$$

In this case the transmission coefficient in Eq. 6 is approximated by $\overline{I/V}$ to remove the distance dependence (division by I) and the voltage dependence (multiplication by V). The overline means that the voltage dependence of I/V is strongly smoothened to avoid singularities in the density of states ρ in the band gap region where no current is measured. Despite these limitations it is well possible to experimentally measure the density of states as a function of the energy relative to the Fermi level using Eq. 8.

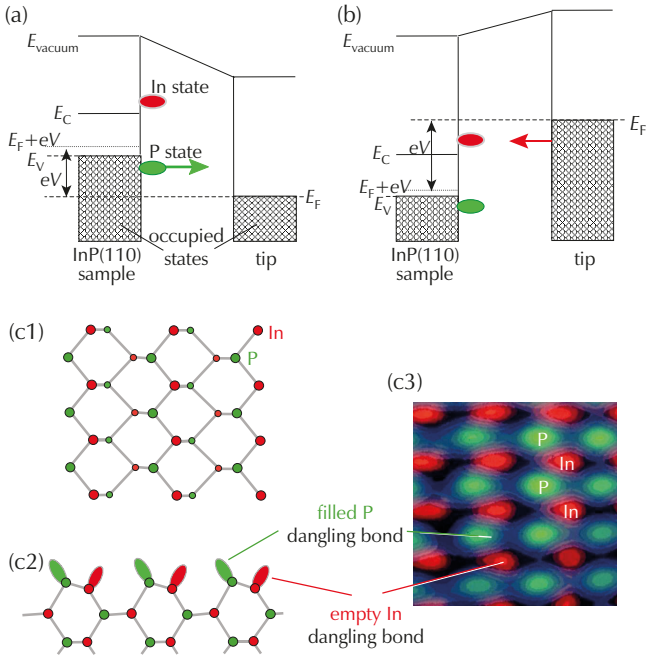


Fig. 6: *Tunneling process at (a) negative and (b) positive voltages applied to the InP(110) surface. At negative voltages occupied states at the P atoms contribute to the tunnel current, while at positive voltages empty In-derived states dominate the current flow. (c1) Schematic top view and (c2) side view of the (110) surfaces of III-V compound semiconductors. (c3) Superposition of two scanning tunneling microscope images measured at positive (red) and negative (green) voltage. The density of state maxima correspond to the surface states at the In and P atoms, respectively.*

2.2 Operating modes

At this stage the experimental operation of a scanning tunneli ng microscope is addressed. The simplest manner to obtain a scanning tunneling microscope image is to directly measure the variation of the tunnel current as a function of the scanning position while keeping the distance z between tip and sample surface constant. A so-called current image is then obtained. Instead of directly recording the atomic variation of the current, however, the usual procedure is to keep the tunnel current constant while scanning over the surface. This is done by changing the distance z between tip and surface using a feedback loop (Fig. 7a). In order to get an image, one records the voltage applied to the piezoelectric crystal (z-piezo), which adjusts the tip-sample distance such that the tunnel current is kept constant as schematically illustrated in Fig. 7b and 7c. This yields a constant-current STM image.

The constant-current mode is the preferred operation mode for most measurements, because it compensates with help of the feedback drifts in the tip-sample separation, which in current STM images would lead to a huge change in current due to the high tip-sample separation

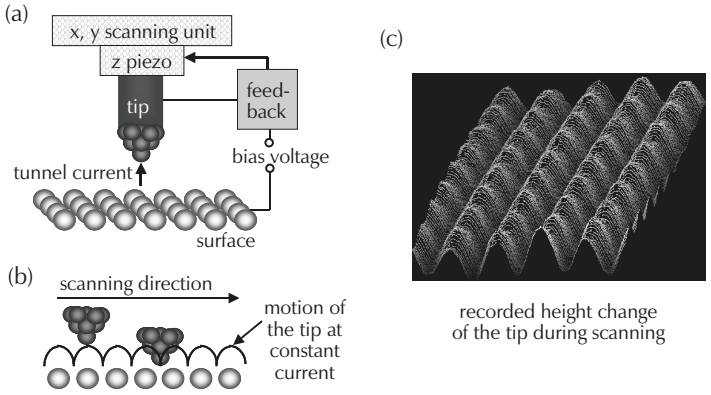


Fig. 7: *(a) Schematic drawing of a scanning tunneling microscope with feedback loop used to keep the tunnel current constant while scanning by adjusting the tip-sample separation using a z piezoelectric element. (b) Motion of the tip in the constant-current mode due to the adjusting of the tip-sample separation. (c) The STM image is obtained by recording the voltage necessary to adjust the tip-sample separation in the constant-current mode for a large number of individual scan lines. The voltage is proportional to changes in the tip-sample separation.*

sensitivity (see Eq. 5). Such drift-induced current changes would otherwise obscure the atomic scale information.

A further operation mode is the spectroscopy acquisition by STM. It is usually done by interrupting the feedback in order to keep the tip-sample separation constant during acquisition of the $I - V$ spectroscopy data. This can be done at any desired surface spot or for every pixel in a STM image. Although the shortly interrupted feedback loop keeps the tip-sample separation in principle constant, there are two effects, which may lead to changes in the tip-sample separation. First a possible drift changing the tip-sample separation can be controlled by a sufficiently long equilibration time of the system and a fast $I - V$ data acquisition. Second, fluctuations in the electronic structure, e.g., dopant atoms or fluctuations of the concentration of dopant atoms, lead to different tip-sample separations for identical set conditions (set current and set voltage). For a proper comparison of spectra from different locations, the individual spectras' tip-sample separations need to be recalibrated by the acquisition of additional current–tip-sample separation curves from which one can determine κ_{eff} in Eq. 5 [25] [26]. Once this is done an exact and quantitative comparison of spectroscopy data from different surface spots is possible.

2.3 Experimental realization

A large variety of different scanning tunneling microscope designs were developed, in order to adjust it best the needs of the individual research projects. Of course, it is not possible to discuss all designs here and it is preferable to refer to selected references [27]- [29]. In the following, a design developed at the Research Center Jülich [30] will be discussed in more detail to outline the general operating principles, which are – with modifications – the same as for other scanning tunneling microscopes. In particular, the surface is always scanned with the aid of piezoelectric adjusting elements.

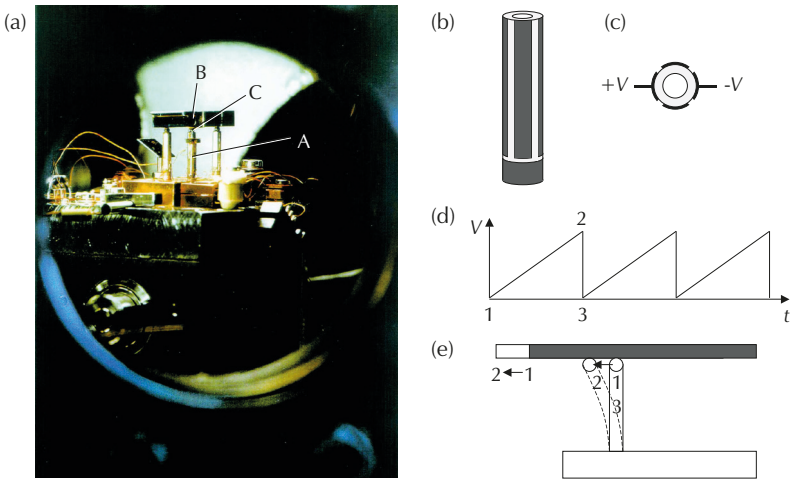


Fig. 8: *Example of a scanning tunneling microscope. (a) View through a window flange at the vacuum chamber showing the STM with the scanner tube (A), the tip (B), and the sample holder (B). (b) Detailed view of a piezoelectric tube with four metallization fields. (c) View from top on a piezoelectric tube showing the metallization fields in dark and the voltage connections for bending the tube. (d) Voltage applied on the piezoelectric tube to shift the sample holder as shown in (e) using its inertia during the fast retreat of the bended tube back to its initial position.*

In order to obtain atomically resolved images, the scanning tunneling microscope must have a high mechanical stability, such that no uncontrolled movements take place between the tip and the sample surface during the measurement. How critical the mechanical design of a microscope is, may be recognized by the fact that the tip must be positioned relative to the sample surface with a precision one order of magnitude better than the measuring accuracy required, i.e. horizontally within approx. 10 pm and vertically within 1 pm. The desired mechanical properties are achieved, for example, with a microscope that consists of radially polarized piezoelectric tubes arranged to form an equilateral triangle (Fig. 8). These three (outer) tubes carry the sample holder with the sample (B in Fig. 8). A fourth tube, the scanning tube, is glued in the center of the triangle. A small z-piezoelectric element, which holds the tip (C in Fig. 8), is mounted on the scanning tube for decoupling the z-motion from the x- and y-scanning motions. The inner metallization of the piezoelectric tubes are electrically connected to ground. Each piezoelectric tube has four additional metallizations on the outside to which, e.g., the scanning voltages or voltages required to move the sample laterally, are applied (Fig. 8b and c). Due to the radial polarization of the piezoelectric material, the tubes can be bent, elongated or shortened.

One of the three outer tubes supporting the sample holder is mounted on a mobile base plate, whereas the other two and the scanning tube are mounted on a common fixed base plate. The coarse approach is achieved by raising and lowering the piezoelectric tube mounted on the mobile base plate. This allows us to adjust mechanically the tip-sample separation until the separation is small enough for the z-piezo's extension to control the adjustment of the tip-sample separation, e.g., keep the tunnel current constant.

The sample holder can be moved by bending the three outer piezoelectric tubes. If one prefers to move the sample over larger distances, then voltage pulses as shown in Fig. 8d are applied. They bend the outer tubes slowly in the desired direction and retract them very fast. Due to the inertia of the sample holder, it can only follow the slow bending, but not the fast retraction (Fig. 8e). Then the piezo tubes will glide back, while the sample holder rests in its bended position. A repetition of this process makes it possible to reach any location on the sample holder.
The whole microscope rests on several damping rings in a ultrahigh vacuum chamber, which is positioned on a compressed-air-damped table of approximately 1 t weight. The measurements are performed at a pressure of less than 1×10^{-8} Pa in the vacuum chamber to ensure that the surfaces remain clean. Fig. 8a shows such a scanning tunneling microscope viewed through one of the window flanges of the vacuum chamber.
The preparation of the tunneling tips is one most crucial part is operating a scanning tunneling microscope, because the tunneling tip as probe is directly affecting the quality of the measurement results. One possibility of preparing tunneling tips is the electrochemical etching of polycrystalline tungsten wire with NaOH. The tips produced in this manner have a radius of curvature of only 5 nm as shown in Fig. 3b and c.

2.4 Applications of the classical scanning tunneling microscope

The scanning tunneling microscope covers a wide field of applications wherever information about the surface structure is required in real space. The applications are so widely distributed over many research fields, ranging from biology to crystallography, that it is essentially impossible to provide a full overview of the possibilities to apply the scanning tunneling microscope. For a more complete overview consult Refs. [29] [31] [32]. Here only selected examples will be presented, which illustrates the potential of a scanning tunneling microscope.

2.4.1. Atomic-scale investigation of surface defects

Due to its high spatial resolution, the scanning tunneling microscope is an ideal instrument for the examination of lattice defects. In particular, the electronic and structural properties of point defects such as individual vacancies and dopant atoms can be measured, which has not yet been possible with other methods on the atomic scale. As an example, we consider P vacancies on InP(110) surfaces. In order to produce a vacancy, an atom must be removed from the surface. Therefore, three bonds are broken and so-called dangling bonds, i.e. unsaturated bonds, are left. These unsaturated bonds do not represent the energetically most favorable configuration and they reconstruct forming three defect energy levels. Defect energy levels are electron states located at or in a vacancy. In the case of high defect concentrations, the defect energy levels change the electrical properties of whole crystals, which is utilized e.g. in doping semiconductor crystals with impurities. As shown in Fig. 6, a scanning tunneling microscope only images either the occupied or the empty states. Since the occupied states correspond to the positions of the phosphorus atoms in the surface, a missing occupied surface state is the signature of a phosphorus vacancy. This missing occupied surface state can be seen in Fig. 9 in the case of phosphorus vacancies on p-doped InP(110) surfaces. In addition, Zn dopant atoms are visible, which are surrounded by a local elevation (Fig. 9) due to their negative charge. In contrast on p-doped InP(110) surfaces the phosphorus vacancies are surrounded by a depression (Fig. 9), because of their positive charge [2] [34].

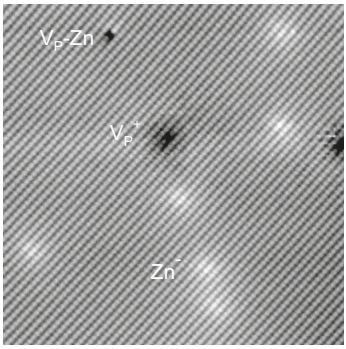


Fig. 9: *Overview of defects occurring on p-doped InP(110) surfaces. In addition to phosphorus vacancies appearing as black depressions (V_P^+, white elevations surounding Zn dopant atoms (Zn^-) can be observed (sample voltage -2.2 V). Adapted with permission from [33], ©1996 American Physical Society.*

2.4.2. Probing the local potential

The above example shows that local screened Coulomb potentials surrounding charged defects and dopant atoms are visible in STM images. The question is now, how can the local potential influence the tunnel current and thus be visible in STM images? Figure 10a shows a STM image of a two-dimensional semiconducting $\sqrt{3} \times \sqrt{3}$ Ga overlayer on Si(111). Each maximum in the empty state STM image corresponds to one empty dangling bond above a Ga adatom. The weaker maxima (marked D) arise from Si atoms located on $\sqrt{3} \times \sqrt{3}$ Ga sites. These Si atoms act as donors and provide the free electrons. The resulting positive charges of the Si dopants induce a redistribution of the free charge carriers and thereby a potential change, which gives rise to the surrounding bright contrast on which the atomic corrugation is superimposed. The local potential change also shows up in the tunneling spectra: the valence E_{V} and conduction band E_{C} edges shift 0.15 eV to higher energies with increasing spatial separation from the dopant site (dotted lines in Fig. 10c) [18].

In addition, the STM images exhibit long-range height changes with lateral extensions of 5 to 15 nm (see dashed and dotted elliptical lines in Fig. 10a). Furthermore, regions with dark and bright contrast appear in surface areas with low and high dopant concentration, respectively. The above observed band edge shifts indicate again that potential changes along the surface give rise to the long-range height changes in STM images.

The sensitivity of the tunnel current to the potential can be illustrated using Fig. 11a, which shows schematically a tunnel contact between a metallic tip and a semiconductor. With no potential change (black lines) the tunnel current is the sum of all electrons tunneling from the electron states between E_F and $E_F + eV$ into the empty sample states. This is schematically shown by the black triangle. In the presence of a band bending (or any potential fluctuation) the band edges is shifted (red lines) and as a result the barrier is modified (red double ended arrow) and the tunnel current increases as shown by the red triangle (in case of a negative potential change). Thereby the tunnel current sensitively changes with any potential fluctuation. The feedback loop keeps the tunnel current nevertheless constant by changing the tip-sample separation, which yields the contrast changes in Fig. 10 a and b.

The underlying potential fluctuations in Fig. 10 a and b can be quantitatively derived from

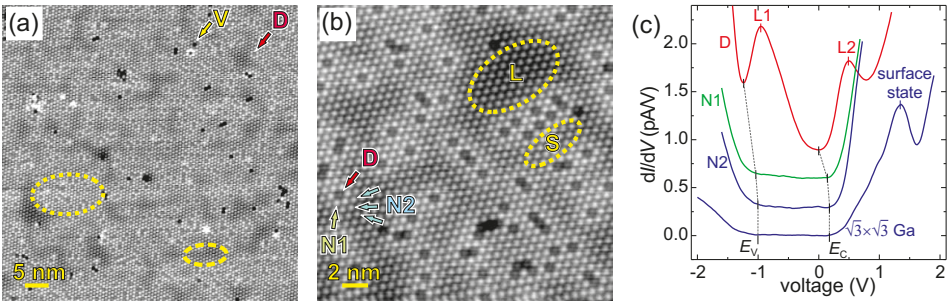


Fig. 10: *(a) STM image of a 2D semiconducting $\sqrt{3} \times \sqrt{3}$ Ga overlayer on Si(111) measured at 2 V. In addition to atomic-sized features arising from vacancies (V) and Si dopant atoms (D), long-range changes in the contrast occur. Examples are indicated by dashed (depression) and dotted (elevation) elliptical lines. (b) High-resolution STM image. Each charged Si dopant is surrounded by a bright contrast. (c) dI /dV tunneling spectra above dopants (D), directly neighboring Ga atoms (N1), Ga atoms further away (N2) [see (b)], and for the $\sqrt{3} \times \sqrt{3}$ Ga overlayer (all Ga sites not neighboring to dopants). For enhanced sensitivity, the spectra D, N1, and N2 were taken at a tip-sample separation 0.06 nm smaller and each curve is offset by 0.3 pA/V for clarity. The conduction (E_C) and valence (E_V) band edges shifts are indicated by a dotted line. L1 and L2 are localized states related to the Si donors. Reprinted with permission from [18].*

the contrast fluctuations by relating quantitatively the *local* height change Δz from the spatial average z_{avg} of tip-sample separation ($z = z_{avg} + \Delta z$) to the *local* potential [18]. This allows to relate the local potential with the local dopant concentration n_{local} as shown in Fig. 11b. This data can be well described [18] by

$$n_{\text{local}} = \frac{6 m_{eff} kT}{\pi \hbar^2} ln\left[1 + e^{-(E_{\text{C,avg}} - E_{\text{F}} + \Delta E_{\text{C}})/kT}\right] \tag{9}$$

as shown by the solid fit curve and the effective masses obtained in Fig. 11c. They correspond well to the effective mass of a Si(111) plane of 0.37 m_0.
This example thus illustrates that it is possible to investigate quantitatively the local potential induced by local fluctuations in the distribution of dopant atoms and/or defects with the aid of a scanning tunneling microscope. The additional deeper analysis performed in Ref. [18] shows that one can even determine the exact origin of the different potential fluctuations in a two-dimensional semiconductor with disordered dopants.

2.4.3. Scanning tunneling microscopy of nano-scale device structures

Using the so-called cross-sectional technique, where a grown sample is cleaved perpendicular to its growth direction to expose a cross-section of the growth structure, it is also possible to investigate hetero- and homostructures with atomic resolution. Figure 12a shows a large scale cross-sectional scanning tunneling microscopy (STM) overview of several 30 nm wide p- and n-doped GaAs layers cleaved along a (110) plane (C and Si dopant atom concentrations of $(5\pm1)\times10^{18}$ and $(4\pm1)\times10^{18}$ cm^3, respectively). An atomically resolved image is shown in

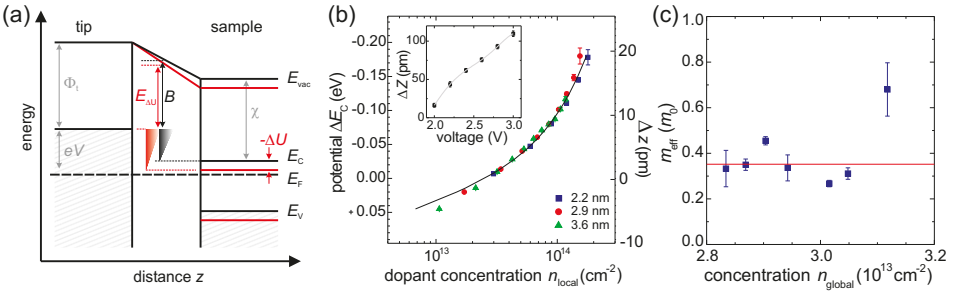


Fig. 11: *(a) Energetic diagram showing the potential sensitivity of the scanning tunneling microscope. The tunnel contact between a metallic tip and a semiconducting surface is shown in black lines. The barrier B is indicated by a black double ended arrow. The total tunnel current is indicated by the black triangle. In case of a potential change in the semiconductor ($-\Delta U$) the barrier changes (red double ended arrow) and the total tunnel current increases in the case of a negative potential change (red triangle). (b) Local deviation from the average potential ΔE_C (left axis) derived from the local height change $-\Delta z$ as a function of the local dopant concentration n_{local} for different resolutions. The solid line is a fit to the data. Inset: corresponding heightvoltage curve at a set current of 0.1 nA, probing the energy (voltage) sensitivity of the tip. (c) Effective mass m_{eff} vs the global dopant concentration n_{global} obtained from fitting different data sets.*

Fig. 12b. The p- and n-doped layers are separated by lines with a darker contrast, whereas the doped layers themselves appear both bright. The n- and p-doped layers were identified on basis of the growth sequence, secondary ion mass spectra, and tunneling spectra showing a typical p and n type behavior.

The dark lines between the p- and n-doped layers were found to be the image of the depletion zones localized at $p-n$ interfaces, where the Fermi-energy is close to midgap [37] [38]. Thus the dark lines mark the *electronic* interface between the n-doped and the p-doped layers. It is important to note that this interface is *not* the atomically sharp metallurgical or chemical interface (marked by dashes), where the doping changes from C_{As} to Si_{Ga} or vice versa. Fig. 12 shows that the electronic interface (dark lines) exhibits rather a roughness much larger than one atomic layer. The white arrows in Fig. 12a point out examples of a local electronically very narrow and wide p-type 'layer'. Note that the individual bright hillocks with about 3 to 5 nm diameter, visible in the p- as well as n-doped layers, are the signatures of negatively charged C_{As} and positively charged Si_{Ga} dopant atoms, respectively [32]. Their contrast is essentially given by the image of the screened Coulomb potential as outlined above [35]. Figure 12b shows that the depletion zone imaged as dark line circumvents each individual dopant atom. Undoubtedly each dopant atom near the metallurgical interface causes a short range meandering of the electronic interface on the scale of 2-5 nm.

Careful inspection of large scale STM images and a quantitative analysis of the interface roughness reveals a further contribution to the interface roughness on a longer length scale, leading to the electronically wide and narrow layers (see examples marked by white arrows in Fig. 12). The physical origin of this second roughness contribution is nicely illustrated in Fig. 12b, which shows that dopant atoms within nominally homogeneously doped layers exhibit large

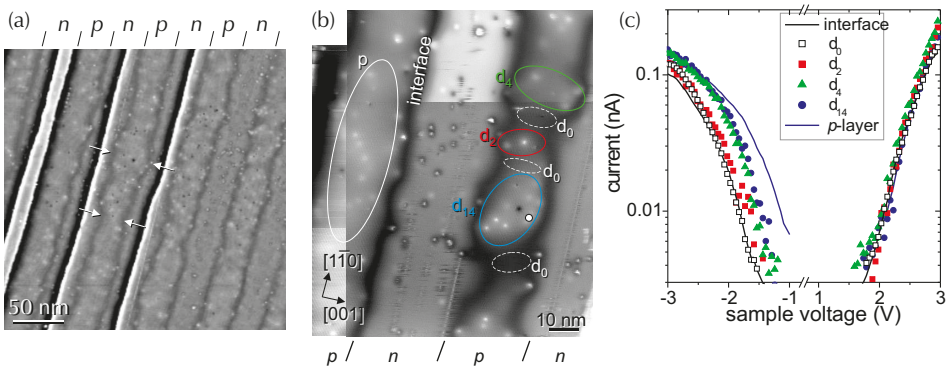


Fig. 12: *Large scale (a) and atomically resolved (b) cross-sectional scanning tunneling microscopy images of multiple p- and n-doped GaAs layers. The bright hillocks marked by Si_{Ga} and C_{As} arise from individual dopant atoms. The dark lines between the p- and n-type layers are the signatures of the depletion zone at each $p-n$ interface and show the position of the electronic interface. Note its pronounced roughness and its correlation with the dopant atoms. The bright hillocks are signatures of dopant atoms. The encircled p-doped areas d_2, d_4, and d_{14} are dopant-induced dots confined by potential barriers due to the doping of the surrounding areas (n-type and lack of dopants (d_0)). (c) Current-voltage curves acquired in these encircled areas in (b). The spectra were normalized to a common tip-sample separation. The growth direction is [001]. Adapted with permission from [25], ©2002 American Physical Society and [36] ©2003 American Institute of Physics.*

variations in the local concentration leading to clusters of dopant atoms as those encircled in Fig. 12b. Above and below are areas locally free of dopant atoms (marked d_0). This clustering not only induces the long range roughness of the electronic interfaces with correlation lengths of about 25 nm and amplitudes of approximately 2.5 nm, but it also drastically modifies the electronic properties.

In order to illustrate this effect the encircled areas, labeled d_2, d_4, and d_{14} according to the number of dopant atoms visible within the area, are of particular interest. These areas are bordered along the growth direction by *n*-doped layers and perpendicular to the growth direction by zones with no dopant atoms (dark contrast areas labeled d_0). These zones with no dopant atoms exhibit tunneling spectra (Fig. 12c) with typical characteristics of a depleted region (as discussed below). Thus the areas (d_2, d_4, and d_{14}) are semiconductor dots, whose confining potential for free holes is defined by the doping of the surrounding and thus by build-in potentials in the order of a few tenths of eV (border toward depleted zones) to 1.4 eV (toward n-doped layer).

Figure 12c shows local tunneling spectra measured above the different cluster areas. In order to allow a proper comparison of the spectra, they were normalized to a common tip-sample separation [38]. All spectra are essentially identical at positive sample voltages. In contrast, at negative sample voltages the current-voltage curves are shifted relative to each other. The fewer dopants are inside the cluster, the greater is the shift towards more negative voltages relative to the spectrum of the spatially extended *p*-doped layer (labeled p in Fig. 1).

For a quantitative discussion of the spectra, we recall that the spectra consist of the current from

valence and into conduction band states at negative and positive voltages, respectively, with the band gap region in between [15] [20]. At a fixed voltage the tunneling current is determined by the energetic positions of the band edges underneath the tip:
At negative voltages, electrons tunnel from all filled valence band states lying between the valence band edge at the surface and the Fermi level of the tip. The size of this energy window is determined by the degree the tip bends the bands at the surface. This tip-induced band bending arises from the fact that the electric field between the tip and the surface penetrates into the semiconductor surface, due to the limited free charge carrier in a semiconductor. Therefore the shifts of the spectra at negative voltages (Figs. 2, 3) indicate that the tip pulls downward the position of the valence band edge at the surface of our dopant-induced dots the more, the less dopants are enclosed within the dot. This suggests a reduced screening ability of the dots with smaller numbers of dopant atoms. This situation is schematically shown in Fig. 13.

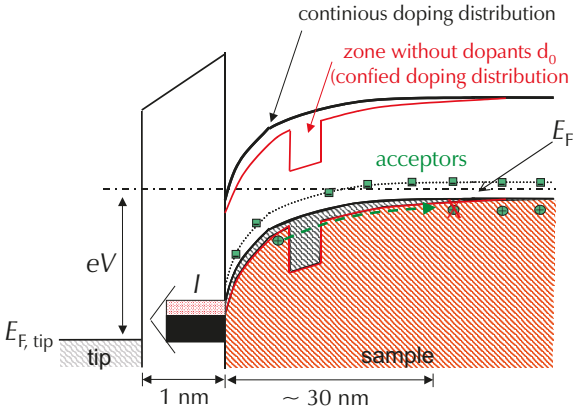


Fig. 13: *Lateral band diagram showing the conduction and valence band edges in the presence of a tip with negative voltages applied to the sample. Two cases are shown. In black: model of continuous doping throughout the sample. In red: inhomogeneous doping with confining barriers arising from areas free of dopants. In that case the depletion zone is extended and therefore the band edge positions at the surface are lowered compared to the unconfined case. As a result the tunnel current is reduced. Adapted from [26].*

The electric field of the tip is screened by negatively charged acceptors, whose free holes have to be pushed away. On this basis, a reduction in the ability to screen the field of the tip is due to a combination of (i) the impossibility to deplete the dopant-induced dots from the free holes and (ii) the number of acceptors within the dot.
(i) The first effect depends on the size of the dot. If the dimension of the dot is much larger than the depletion width induced by the screening of the tip's field, i.e., the field can be screened entirely by the acceptors within the dot, then the band bending is determined by the concentration of dopant atoms in the dot (band edges drawn in black lines in Fig. 13). This effect is applicable to the spatially extended p-doped layer on the left hand side of Fig. 12a, which is electronically homogeneous over dimensions of more than 100 nm, which is much larger than the estimated depletion width of 10 to 25 nm for acceptor concentrations of $4 \times 10^{18} cm^{-3}$ at -2.5 V sample voltage.

If the dots' dimensions are similar to the width of the depletion zone, one also needs to consider whether the holes can actually be pushed away sufficiently to accommodate the screening of the tip's electric field. For the small dot sizes, this means that holes need to be pushed out of the enclosed dot areas to deplete them . We recall that our dots are confined by potential wells. These wells act as a barrier for the free holes and impedes the holes to be pushed out (see band edges drawn in red lines in Fig. 13), such that the remaining negatively charged acceptors could screen the electric field of the tip. If the dot cannot be sufficiently depleted, the band bending increases, causing lower tunneling currents than expected for infinitely sized bulk GaAs crystals with the same dopant concentration. The effect of confinement of the free holes on the tunneling current becomes smaller the higher the applied voltage, because the relative fraction of states inhibited to tunnel by the locally increased tip-induced band bending diminishes as more valence band states are involved in the tunneling process.

At positive sample voltage no dependence of the current on the number of enclosed dopant atoms is found, implying that the energetic position of the conduction band edge underneath the tip is the same for all areas investigated here. At positive sample voltage the electric field of the tip is screened by free holes accumulating at the *p*-doped surface. These holes feel no barrier to accumulate at the surface and thus can effectively screen the tip's field. Furthermore, the number of accumulated holes is determined by the density of valence band states, which is the same for all investigated areas, because the material is the same. This explains the almost invariant current-voltage spectra obtained on the various dots at positive sample voltages.

(ii) The above discussion also shows, that the ability to screen the electric field of the tip is directly proportional to the number of acceptors available within the electronically isolated dot. Thus the fewer acceptors within the dot, the larger the tip-induced band bending and the larger the voltage shift observed in the tunneling spectra.

These results show that local variations in the dopant atom distribution lead in nanoscale semiconductor structures to uncontrolled Fermi-energy positions in space and energy, effectively limiting the miniaturization of semiconductor devices.

3 The scanning force microscope

The design and development of the scanning force microscope (SFM) (frequently called atomic force microscope AFM) is very closely connected with that of scanning tunneling microscopy. The central component of both types of microscopes is essentially the same. A fine tip is positioned at a characteristic small distance from a sample. The height of the tip above the sample is again adjusted by piezoelectric elements. The images are measured by scanning the sample relative to the probing tip and measuring the interaction of the tip with the sample. However, a scanning force microscopy is now not measuring the tunnel current, but rather the forces between the tip and the sample. This is achieved by mounting the tip on a spring blade, called cantilever, whose deflection is probed as a function of lateral position.

Initially, the height deflection was measured by a scanning tunnelling microscope piggybacked on the spring. However, this method is very demanding. Nowadays, other optical techniques are preferred for measuring the deflection. A rich variety of forces can be sensed by scanning force microscopy. In the non-contact mode (of distances greater than 1 nm between the tip and the sample surface), van der Waals, electrostatic, magnetic or capillary forces produce images, whereas in the contact mode, ionic repulsion forces take the leading role. Because its operation does not require a current between the sample surface and the tip, the SFM can be used to

investigate non-conductive materials inaccessible to the scanning tunneling microscope. For example, insulators, organic materials, biological macromolecules, polymers, ceramics, and glasses are some of the many materials which can be imaged in different environments, such as in liquids, under vacuum, and at low temperatures using the SFM.
In the non-contact mode one can obtain a surface analysis with a true atomic resolution. However, in this case the sample has to be prepared under ultrahigh vacuum conditions. The non-contact mode has the further advantage over the contact mode that very soft and/or rough surfaces are not influenced by frictional and adhesive forces between the tip and the sample during the scanning process, i.e. the surfaces are not mechanically scratched.

3.1 Theoretical principles of the scanning force microscope

As already mentioned above, van der Waals forces lead to an attractive interaction between the tip on the spring and the sample surface. The potential can be described in a simpler classical picture as the interaction potential between the time dependent dipole moments of the two atoms. Although the centers of gravity of the electronic charge density and the charge of nucleus are exactly overlapping on a time average, the separation of the centers of gravity is spatially fluctuating in every moment. This produces statistical fluctuations of the atoms' dipole moments. The dipole moment of an atom can again induce a dipole moment in the neighboring atom and the induced dipole moment acts back on the first atom. This creates a dipole-dipole interaction on basis of the fluctuating dipole moments. This attractive interaction decreases with z^{-6}, z being the separation of the two atoms. At very small distances, the interaction is dominated by the repulsive interaction between the atomic cores, which is decreasing $\propto z^{-12}$. The attractive van der Waals potential and the repulsive core interactions both together form the Lenard-Jones potential.
At larger distances, the van der Waals interaction potential decreases more rapidly (z^{-7} instead of z^{-6}). This arises from the fact that the interaction between dipole moments occurs through the exchange of virtual photons. If the transit time of the virtual photon between atoms 1 and 2 is longer than the typical fluctuation time of the instantaneous dipole moment, the virtual photon weakens the interaction. This range of the van der Waals interaction is therefore called retarded, whereas that at short distances is unretarded.
The scanning force microscope is not based on the interaction of individual atoms only. Both the sample and the tip are large in comparison to the distance. In order to obtain their interaction, all forces between the atoms of both bodies need to be integrated. The result of this is known for simple bodies and geometries. In all cases, the summation leads to a weaker decrease of the interaction. A single atom at distance z relative to a half-space leads to an interaction potential of

$$U = -\frac{C\pi\rho}{6} \cdot \frac{1}{z^3} \tag{10}$$

where C is the interaction constant of the van der Waals potential and ρ the density of the solid. C is essentially determined by the electronic polarizabilities of the atomic species of the bodies (or atoms) 1 and 2. If one has two spheres with radii R_1 and R_2 at distance d (distance between sphere surfaces) one obtains an interaction potential of

$$U = -\frac{AR_1R_2}{6(R_1 + R_2)} \cdot \frac{1}{z} \tag{11}$$

where A is the so-called Hamaker constant. It is materials specific and essentially contains the densities of the two bodies and the interaction constant C of the van der Waals potential. If a sphere with radius R has a distance z from a half-space, an interaction potential of

$$U = -\frac{AR}{6} \cdot \frac{1}{z} \tag{12}$$

is obtained from Eq. (11). This case describes best the geometry in a scanning force microscope and is widely used. The distance dependence of the van der Waals potential thus obtained is used in analogy to the distance dependence of the tunnel current in a scanning tunneling microscope to achieve a high resolution of the scanning force microscope. However, the distance dependence being much weaker, the sensitivity of the scanning force microscope is lower.

3.2 Operation principle of scanning force microscope

Two operation modes, the contact and the non-contact modes, are primarily used. In the contact mode the cantilever's tip touches the surface and follows mechanically the topography while scanning. Due to the large tip-sample interactions, this method may modify the surface and thus cannot reliably achieve atomic resolution. Therefore the non-contact mode is preferred for higher resolution. Here again, one has two possibilities of operation. The first one, the static SFM mode probes the force exerted on the tip by the sample and the feedback keeps this force constant, by varying the tip-sample separation. This method has, however, as main disadvantage that it is difficult to implement [54]. The interpretation of the images is, however, straightforward. The image measures a surface of constant force.

In contrast, the dynamic operation method of a scanning force microscope has proved to be particularly useful and widely applicable. In this method the nominal force constant of the van der Waals potential, i.e. the second derivative of the potential, is exploited. This can be measured by using a vibrating tip usually at frequencies in the range of 50 to 500 kHz (Fig. 14a). If a tip vibrates at distance z_∞, which is outside the interaction range of the van der Waals potential, then the vibration frequency and the amplitude are only determined by the nominal force constant k of the spring (Fig. 14b). This corresponds to a harmonic potential. When the tip is moved into the interaction range of the van der Waals potential, the harmonic potential and the interaction potential are superimposed thus changing the vibration frequency and the amplitude of the spring.

This is described by modifying the nominal force constant k of the spring by an additional fraction f of the van der Waals potential. As a consequence, the resonance frequency of the cantilever is shifted to lower frequencies as shown in Fig. 14c. ω_0 is the resonance frequency without interaction and $\Delta\omega$ the frequency shift induced by the presence of the tip-sample interaction. If a constant excitation frequency of the tip ω_m is selected such that $\omega_m > \omega_0$, the amplitude of the vibration decreases as the tip approaches the sample, since the interaction becomes increasingly stronger. Thus, the vibration amplitude also becomes a measure for the distance of the tip from the sample surface. If a spring with low damping Q^{-1} is selected, the resonance curve is steep and the ratio of the amplitude change for a given frequency shift becomes large.

Typically small amplitudes (approx. 1 nm) in comparison to tip-sample distance z_m are used to ensure the linearity of the amplitude signal. With a given measurement accuracy of 1 %, however, this means that the assembly must measure deflection changes of 0.01 nm, which is achieved most simply by a laser interferometer or optical lever method.

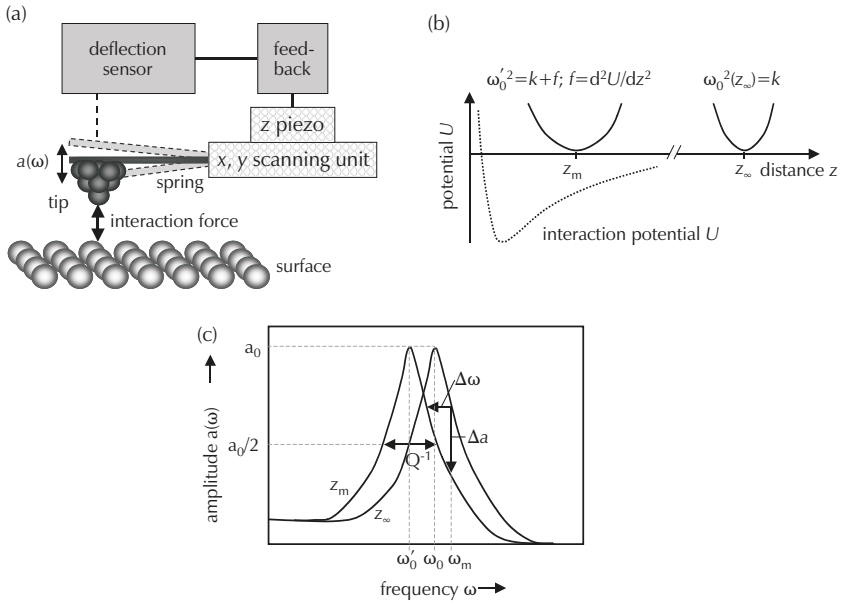


Fig. 14: *(a) Schematic of the principle of operation of the dynamical scanning force microscopy mode. (b) Schematic illustration of the effect of the van der Waals potential on the vibration frequency of the tip. On the right side the tip is far away from the surface (resonance frequency ω_0 at distance z_∞) and on the left side close to the surface (distance z_m) within the extension of the van der Waals potential. The resonance frequency ω'_0 is shifted by the presence of the van der Waals potential. (c) Schematic illustration of the resonance curves of the tip with and without interaction with the sample. ω_m is the vibration frequency of the tip during measurement. The presence of the interaction potential shifts the resonance curves (frequency) and thus at constant tip vibration frequency ω_m the vibration amplitude is changes as a function of the tip-sample separation. Adapted from [49].*

During the measurement, the vibration amplitude and frequency of the tip is measured (see for the measurement principle Section 3.3). A feedback electronics keeps then the amplitude of vibration constant, by changing the tip-sample separation (see Fig. 14a). Thereby the resonance frequency of the cantilever-sample system is kept constant, which implies that the force constant introduced by the tip-sample interaction remains unchanged. Thus, in the dynamic non-contact operation mode the scanning force microscopy measures a surface of constant effective force constant. It does not probe the force or potential itself. This fact needs to be kept in mind, when interpreting SFM images, especially at interfaces and defects. If a sample is chemically homogeneous and only van der Waals forces act on the tip, then the SFM measures the topography of the surface.

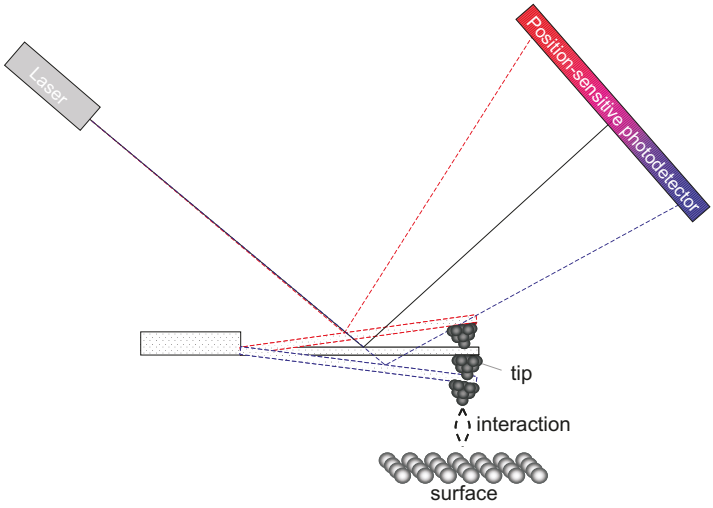


Fig. 15: *Simplified schematic illustration of the optical measurement of the cantilever deflection by laser beam reflection.*

3.3 Technical realization of a scanning force microscope

The main electronic components of the SFM are the same as for the STM, with one exception: Instead of measuring the tunnel current, the topography of the scanned surface is reconstructed by analyzing the deflection of the tip at the end of a spring. The deflection is typically probed by an interferometrical and optical lever method. A common method for probing the deflection of the cantilever is the measurement of the position of a laser-beam reflected from the cantilever on a position-sensitive photo-detector. The principle of this optical lever method is presented schematically in Fig. [**?**] using a position sensitive photo-detector, e.g., one with four separate quadrants. The beam is aligned such that without a displacement of the cantilever all quadrants of the photodiode have the same irradiation. Once the cantilever is displaced, the reflected laser beam moves on the photodiode and the amount of light shining on each of the quadrants is different. Thus the relative signals of every quadrant provide a measure of the defection. Note the precision of this measurement is very high due to the large separation between the cantilever and the position-sensitive photo-detector. The measurement principle and accuracy will be elaborated in more details in Chapter C8, Piezoelectric atomic force microscopy. An overview of the measurement techniques for the deflection can be found, e.g., in [51].
At this stage we have to address the probe, i.e. the tip mounted on the spring, which is the core of the whole microscope. The first cantilevers were made of a gold foil to which a diamond tip has been attachted [4]. Nowadays commercial cantilevers are mostly etched from single crystalline silicon wafers using the technology applied to the manufacturing of integrated circuits [52]. The resulting cantilevers have a typical geometry as shown in Fig. 16. The cantilever is characterized by a spring constant k, an eigenfrequency ω_0, and a quality factor Q. The spring constant for the geometry shown in Fig. 16 is given by [53]

$$k = \frac{E_Y w t^3}{4L^3} \tag{13}$$

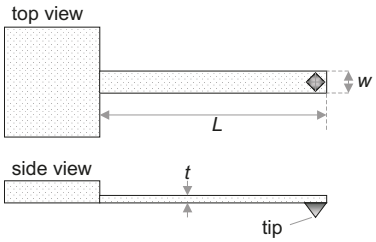


Fig. 16: *Schematic top and side view a cantilever as typically used for scanning force microscopy*

with w being the width, t the thickness, and L the length of the cantilever. E_Y is Young's modulus of the cantilever material. The eigenfrequency of the cantilever is given by [53]:

$$\omega_0 = \frac{0.162t}{2\pi L^2}\sqrt{\frac{E_Y}{\rho}} \tag{14}$$

with ρ being the mass density of the cantilever material.

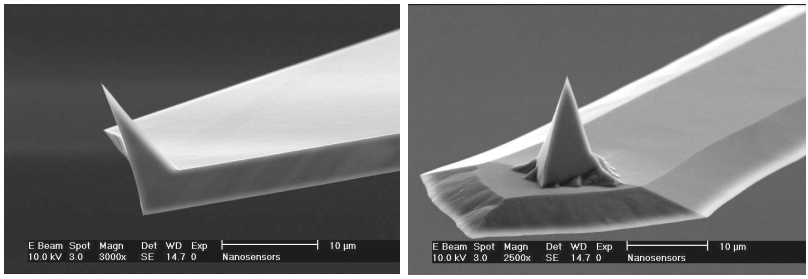


Fig. 17: *Scanning electron microscope images of two commercial cantilevers with different tip geometries. Images courtesy of Nanosensors*

Depending on the application, the key parameters of the cantilever (k, ω_0, and Q) can to be tuned for optimal performance of the SFM by changing the dimensions and the material. One example of a real cantilever produced from a Si wafer is shown in Fig. 17.

3.4 Applications of the scanning force microscope

To a large degree the scanning force microscope is applied to determine routinely topographic information of surface structures. The advantage of the scanning force microscope lies, however, in the wide applicability by changing the interacting force probed. One of the most prominent example is the use of magnetic interaction in the scanning force microscope. In the following, Two applications of the scanning force microscope will be presented, which surely cannot cover the full width of applications (for that see Ref. [29]), but provide a first insight into (i) the use of magnetic forces to probe magnetic structures, and (ii) the electrostatic scanning force microscopy. Note, the SFM is the scanning probe microscope with the widest application in industry. A overview of industrial applications is given in Ref. [55]

4.4.2 Magnetic Scanning Force Microscopy (MFM)

Thus far, we only considered van der Waals forces acting between the tip and the sample. There are, however, other tip-sample forces which can be used in the SFM. For example, if a magnetic tip is mounted on the cantilever, magnetic interactions between the tip and sample can be used to image magnetic surface structures. Magnetic scanning force microscopy is of particular interest for the investigation of magnetic storage media and related nano-magnetic phenomena in thin magnetic films and structures.
In the most general case, the magnetic force between the sample and the tip is given by

$$F_{mag} = -\nabla \int_{tip} \boldsymbol{M}_{tip} \cdot \boldsymbol{H}_{sample} dV \tag{15}$$

or

$$F_{mag} = (\boldsymbol{m}_{tip} \nabla) \boldsymbol{B}_{sample} \tag{16}$$

with $\boldsymbol{H}_{sample.}$ and $\boldsymbol{B}_{sample}$ being the magnetic stray field and the magnetic induction of the sample, respectively. $\boldsymbol{M}_{tip}$ and $\boldsymbol{m}_{tip}$ are the magnetization and the magnetic moment of the tip, respectively. Since in most cases the exact magnetic structure of the tip is not known, a model tip magnetization must assumed. In the simplest case, the tip is a spherically structured magnetic single domain with the magnetization $\boldsymbol{M}_{tip}$.

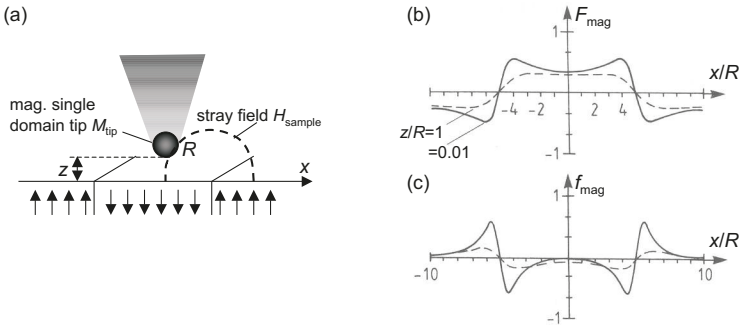


Fig. 18: *(a) Schematic of a sample with magnetic domains investigated by a magnetic tip with a magnetic domain of radius R at a distance z from the sample. (b) Schematic diagram of the spatial variation of the magnetic force F_{mag} for two different tip-sample separations. The length scales were all normalized by the tip radius R. (c) same as (b) but for the magnetic spring constant f_{mag}. The SFM images the spring constant. Thus, a contrast in the magnetic SFM images is found primarily at domain boundaries. Adapted from [57].*

Of particular interest are the stray fields of magnetic storage media which consist of different domains. Since the important aspect in force microscopy is not the forces but the force gradient, i.e. the force constant f, a pronounced variation of the signal is found near the domain walls, but not inside a domain. This situation is sketched in Fig. 18. The parameter of the two curves shown (solid and broken lines) is the ratio of the working distance z and the radius R of the magnetic domain of the tip.
Fig. 19 shows an experimentally measured picture of four different oriented magnetic domains. Images **b** and **c** show the fine structure of a 180° domain wall. Alternating bright and dark

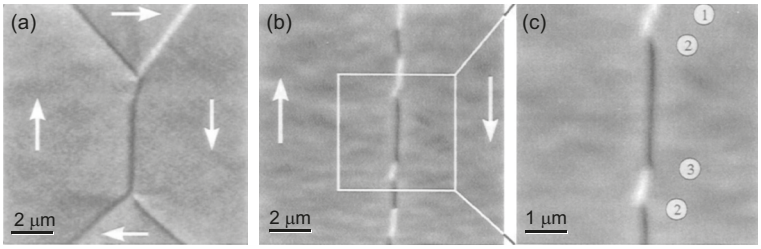


Fig. 19: *(a) Magnetic SFM image of a Landau-Lifshits domain structure in a 500 nm thick Fe film on a 150 nm Ag/ 1 ML Fe/ GaAs(100). (b) Magnetic SFM image of a 180 ° domain wall, ehibiting areas of opposite magnetic orientation. (c) Zoom-in of (b). In all cases the domain walls dominate the contrast in magnetic SFM images. Adapted from [58].*

contrasts can be seen. These contrast changes show that the domain wall consists of segments with different wall orientation. This example illustrates that magnetic SFM is well suited for imaging magnetic structures that are commonly used in today's' storage media.

4.4.3 Electrostatic Scanning Force Microscopy (EFM)

The electrostatic scanning force microscopy is a SFM technique, based on the electrostatic Coulomb interaction. The EFM characterizes thus the electrical properties, i.e. the charge distribution of a sample. For the realization of the EFM the tip must be insulated.
When a tip on a cantilever is approached close to a sample and a voltage V is applied, then the EFM forms a capacitor, having a complex geometry. The Coulomb interaction between the tip on a cantilever and the sample is given in such a case by [59]:

$$F_{el} = \pi\epsilon_0 V^2 \cdot g(z) \tag{17}$$

where $g(z)$ is related to the change of capacitance with changing the tip-sample separation z, i.e. $g(z) \propto dC/dz$. The change of capacitance with varying z is a function of the tip geometry (e.g., tip radius R, angle θ, and height above the cantilever) and the cantilever dimensions length, width, and inclination angle.
For a tip cone with spherical apex, g is given by $-R/z$, hence

$$F_{el} = -\pi\epsilon_0 V^2 \frac{R}{z} \tag{18}$$

for $z/R \ll 1$. For other tip geometries $g(z)$ is much more complicated. Here we concentrate on the simple geometry.
The EFM will in a non-contact mode probe a surface of constant $\partial F_{el}/\partial z$. In general, the extraction of the charge distribution on a surface using the EFM is not straightforward, because the functional dependencies of the Force on the charge distribution are rather complicated and change with different tip geometries.
Therefore, a modern version of electric force microscopes uses for the determination of surface charge or surface potential a complicated lock-in and phase loops electronics (see e.g., [60]). The essential modification with respect to the old apparatus is the so called two-pass technique (LiftMode). In this method each line must be scanned twice. On the basis of two line scans, in which the first represents the topography of the surface in a contact mode and the second is

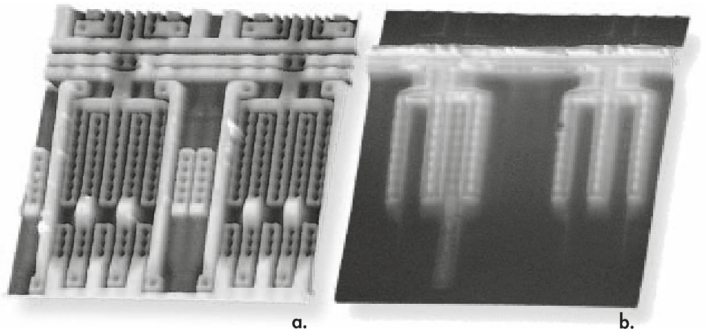


Fig. 20: *(a) Topography and (b) electrostatic scanning force microscope (EFM) image of a live (voltage applied) packaged IC with a passivation layer on. The EFM image detects a transistor in saturation. Image size: 80 × 80 μm^2. Image courtesy of Veeco Instruments Inc. [60].*

taken at a fixed distance relative to the surface, one can reconstruct precisely the distribution of the charge or the potential on the surface without topographical error. In Fig. 20 an example case is given, which highlights the possibilities of this modern tool for microelectronics.

4 Application of scanning probe microscopy in memristive cells

The scanning probe microscopy represents a powerful tool for the investigation of memristive cells like they are used e.g. in ReRAM devices. Although the switching process itself should be considered as a bulk process, the purely surface sensitive analyze technique can be used in many cases for the elucidation of the underlying mechanism on the nanoscale.

A simple ReRAM cell consists of a redox active layer, which is connected to a bottom and a top electrode. By applying an electric potential between the electrodes this layer, which may be e.g. a transition metal oxide layer, can switch between different redox states and thereby change its electronic conductance. In many cases this is realized by a nanometer sized conductive filament, which has been formed by a so called electroforming step. Hereby, a virgin cell is treated once with a significantly higher forming voltage, which leads to the formation of a strongly conductive channel through the switching layer. Depending on the used material systems the reason for the higher conductivity of such a filament can be either a strong increase of oxygen vacancies compared to the surrounding or an enrichment of metal atoms.

In a typical scanning probe experiment for the investigation of resistive switching phenomena one of the two electrodes is replaced by a metallic STM tip or an electronically conducting cantilever. In case of a cantilever the potential is directly applied in contact mode by a gap voltage between the tip and the sample, whereas in STM mode the bias voltage is used to control the switching of the resistive layer. Depending on the potential height either the write or the readout process of a ReRAM cell can be simulated. Therefore, typical values of 2–5 V (write) or 100–500 mV (readout) are used. It should be remarked here, that due to the small radius of such a nanotip often in combination with adsorbate layers on the sample the needed minimum potentials for these processes are in many cases higher than in real devices.

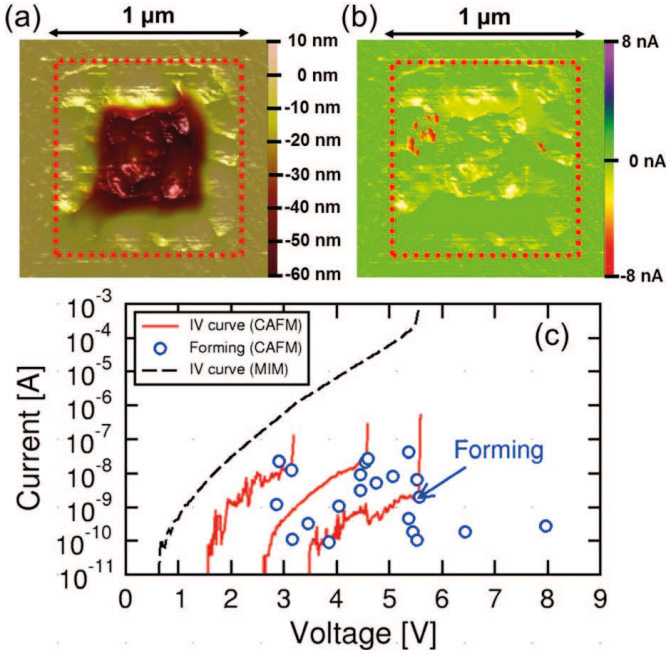


Fig. 21: *(a) Topography map obtained by LC-AFM scanning of the NiO surface, evidencing the enhanced roughness in the plug region due to W dishing during the CMP step following the metal filling of the via hole. (b) Topography/current map after forming, showing surface topography with a light/dark scale and low/high currents as green/red color scale. Red spots reveal electrically-formed conductive filaments at the nanoscale. (c) Measured I-V characteristics for LC-AFM forming (red lines) and scatter plot of forming points at V_{form}, I_{form} (blue circles), compared with I-V characteristic of large-scale MIM device (black dashed line) [61].*

4.1 Local conductivity atomic force microscopy (LC-AFM)

For the detection and manipulation of the electric properties of a memristive layer the local conductivity atomic force microscopy (LC-AFM) is the most common technique. This is mainly attributed to the experimentally rather simple setup and the fact, that the applicability of the method is independent from the electric properties of the sample. However, due to the physical contact of the cantilever tip a mechanical damage of the sample surface (especially in case of soft materials), represents a not to be neglected factor, which can be minimized by carefully tuning the force constant. Another challenge is often the realization of a stable current flow between the tip and the sample because the detected current can be strongly influenced by the collection of surface adsorbates or sample material with the tip during the scanning process.

In general two different kinds of measurements can be realized with the LC-AFM technique. By using a stationary tip position and ramping the gap voltage, I-V measurements can be performed, which leads to typical conductance curves as known from macroscopic device measurements.

Fig. 21 shows the basic principle by switching a 15 nm thick NiO layer deposited on a vertical

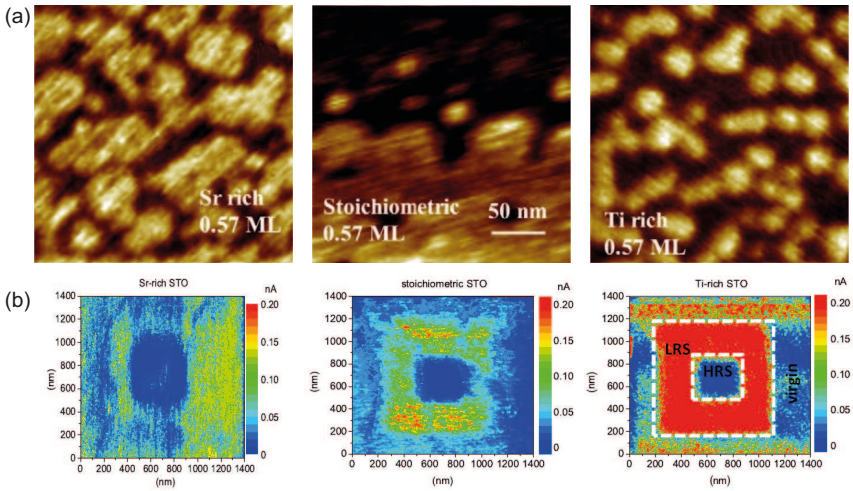


Fig. 22: *(a) c-AFM images (200 nm × 200 nm) of 0.6 ML thick STO films formed under different growth conditions and (b) LC-AFM images of 16 ML thick STO films with different stoichiometries. The low resistive state area was inscribed with $V_{set} = +1.3$ V, the high resistive state area was inscribed into the low resistive state area with $V_{reset} = -1.3$ V. The readout voltage V_{read} is $+0.2$ V [62].*

W bottom electrode, which is surrounded by a 300 nm thick insulating SiO_2/Si_3N_4 stack [61]. The central 0.6 × 0.6 μm^2 depressed region in the topography image (Fig. 21a), lying 20–30 nm below the surrounding area, is due to the dishing of the W-plug during the preparation process. By applying single voltage sweeps over randomly chosen spots within the W-plug area conductive filaments can be visualized in the current map (Fig. 21b).

Another typical experiment is the lateral switching of a sample with the tip. Scanning the surface with an applied gap voltage between the tip and the sample enables the modification of the resistivity of a larger surface area. By using a lower readout voltage this modification can be visualized in a current map.

Fig. 22 shows lateral switching of submonolayer thick $SrTiO_3$ films, which had been grown by pulse laser deposition on Nb-doped $SrTiO_3$ single crystals [62]. By varying the growth parameters the stoichiometry of the films can be controlled, which has a direct influence on both the topography and the resistive switching properties.

The LC-AFM technique has also been used to demonstrate the effects of switching and electroforming of real devices. Therefore, a ReRAM cell is formed and switched to different resistance states in an initial process. Afterwards, the top electrode is removed by a delamination process before the so prepared cell is then transferred into the AFM chamber. Fig. 23 and 24 show the results of such an experiment performed on a 30 nm thick TiO_2 film, which had been covered by a Pt top electrode [63]. Electroforming results in the creation of localized conductance channels induced by oxygen evolution (Fig. 23), while subsequent resistive switching causes an additional conducting structure next to the forming spot (Fig. 24). The lateral extent of this structure depends on the number of switching cycles indicating an ongoing breaking of existing and creation of neighboring current channels during subsequent switching.

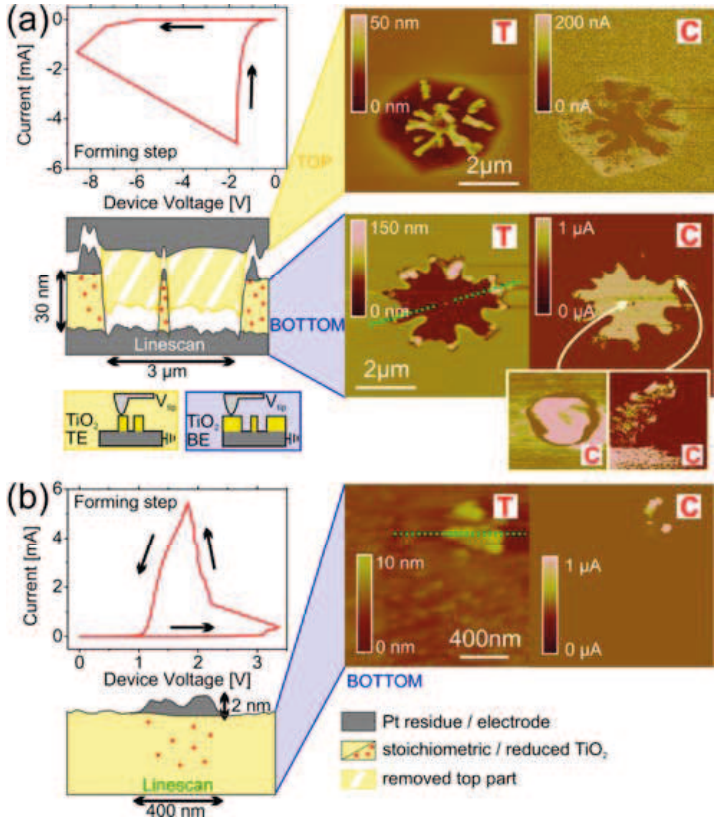


Fig. 23: *Influence of electroforming on the morphology and local conductivity of a sample. The I-V and LC-AFM data of a sample formed by a negative (positive) voltage sweep are shown in part (a) (part (b)). The dashed green line marks the position of the line scan shown on the left hand side [63].*

Investigating the resistive switching properties on the nanoscale with scanning probe methods can give insights into mechanistic details, which are not easily accessible with other methods. E.g. it could be shown, that switching $SrTiO_3$-based memristors is not only realized via conductive filaments but also by a homogenous mechanism [64]. As shown in Fig. 25 both types of switching occur at different threshold voltages but beneath the same electrode. Interesting is that they exhibit the opposite switching polarity, which may be of importance for understanding the behavior of real devices.

The above mentioned impact of an AFM-tip in contact mode on the surface morphology can also be used for a very interesting experimental approach. Through selective ablation of the top surface layers of a switching layer by using high force constants during scanning it is possible to realize a 3D tomography image of a conducting filament [65]. Usually, very hard cantilever tips with sufficient electrical conductance like boron-doped diamond tips are used for such kind of experiments. The obtained 2D current images at different sample depths can then be put together by software to form a 3D image of the filament. With this method it could be

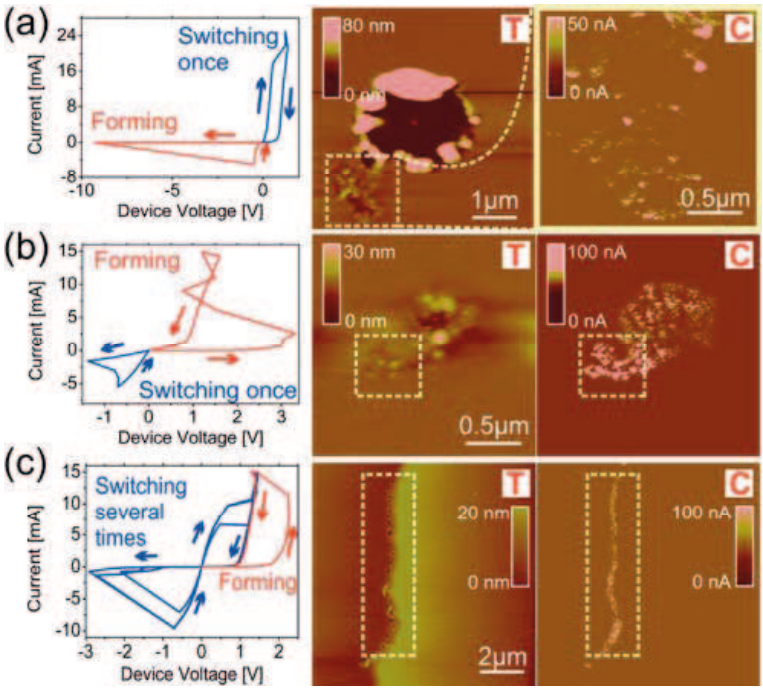


Fig. 24: *Influence of resistive switching on the morphology and local conductivity of the sample. I-V and LC-AFM data are shown for a sample negatively formed and switched once (a), a sample positively formed and switched once (b), and a sample positively formed and switched several times (c). In addition to the forming spot further conducting areas emerge due to the switching steps. They are marked by dashed rectangles [63].*

shown that a filament formed within a 5 nm thick amorphous Al_2O_3 layer sandwiched between a 10 nm TiN bottom electrode and a 40 nm thick Cu top electrode exhibits a conical shape with the narrow part close to the inert electrode (Fig. 26). On the basis of this shape, it was concluded that the dynamic filament growth is limited by the cation transport in that material system.

4.2 Scanning tunneling microscopy

Using the electric field of a metallic STM tip in tunneling mode for resistive switching experiments has the advantage, that the tip is never touching the surface during the measurements, which excludes the possibility of a surface modification by contact effects. The modification of the redox state and, thus, the resistance of the sample are caused by the strong electric field under the STM tip at high bias voltages. A unique advantage of the STM technique is the possibility to obtain information about the local DOS structure of a switched surface layer via STS measurements, which allows a better insight into the underlying mechanism. Furthermore, in contrast to LC-AFM the lateral resolution is not limited by the cantilever dimensions but can reach even atomic resolution.

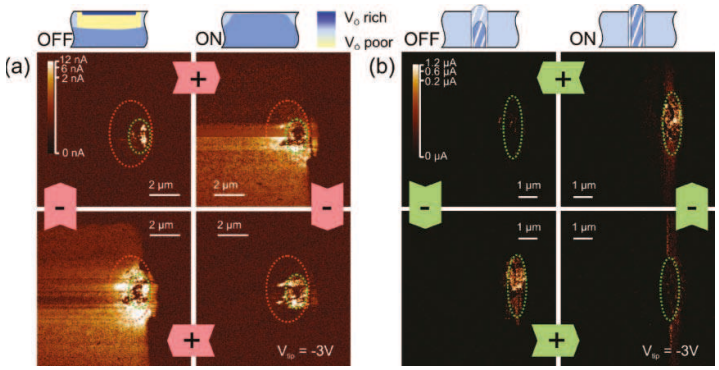


Fig. 25: *Tip induced resistive switching of a delaminated sample. (a) A positive voltage of +5 V will switch the broad area around the crater from Off to On state, while a negative voltage of −5 V switches it back from On to Off state. (b) Going to a higher current compliance allows resolving the crater itself. It can be switched as well (using tip voltages of +/−6 V), but exhibits the opposite switching polarity [64].*

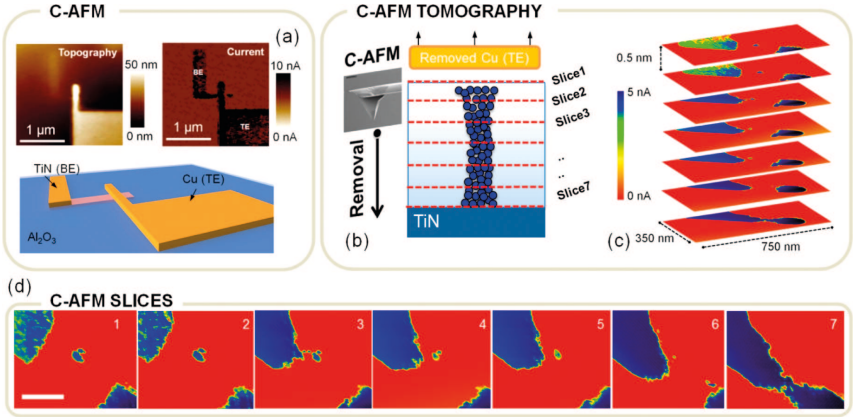


Fig. 26: *(a) Planar 2D LC-AFM, performed on a Cu/Al_2O_3/TiN cross-point memory element (bottom inset) in SET-state. The LC-AFM is completely ineffective due to the presence of the top electrode shielding the conductive filament observation. (b) Schematic of the LC-AFM tomography procedure, the diamond tip is exploited to collect several slices at different heights of the conductive filament after the removal of the top electrode. (c) Over imposition of the collected 2D LC-AFM slices, prior to the 3D interpolation. Note, the average space between each slice is ∼0.5 nm. (d) Collection of 2D slices constituting the data set for the 3D interpolation (scale bar 80 nm). The conductive filament appears in the middle of the active area after top electrode removal. The highly conductive features on the top-left and bottom-right corners are the exposed parts of the TiN bottom electrode, which is progressively exposed during the removal of Al_2O_3 [65].*

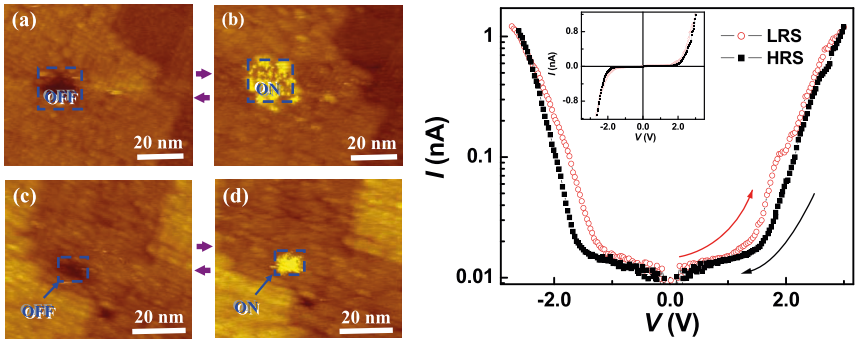


Fig. 27: *STM images (set point: 0.1 nA at 1.8 V) of a Nb-doped $SrTiO_3$ single crystal surface, which characterize reversible surface modifications via the electric field of a STM tip. The OFF and ON areas of blue dashed square shown in these images were obtained by processing this area with writing scan (set point: 0.5 nA at $V \geq 2.0$ V) and reversal scan (set point: 0.5 nA at $V \leq -2.0$ V), respectively. Tunneling I-V curves drawn in semilogarithmic and (inset) linear current scale indicate switching between different resistive states. The directions of voltage scans are indicated by arrows [66].*

Although the application of STM in resistive switching studies is a rather new field of research some nice results can already be found in the literature. Both lateral switching by scanning a surface area with increased bias voltage and single point switching by performing I-V spectroscopy at a constant tip position have first been shown on a Nb-doped $SrTiO_3$ single crystal surface [66]. Fig. 27 shows the field induced change of the image contrast indicating a change of the electronic states density caused by different redox states of the surface atoms. Resistive switching is also indicated by the typical hysteresis curve in the tunneling spectra. Furthermore, a feature near to the conduction band in the LRS state can be attributed to a donor-like level, which agrees to the common mechanism of resistive switching in perowskite materials via oxygen vacancy movement.

The usability of STM for switching or device characterization has been largely restricted to samples with a sufficiently high electronic conductivity, e.g. Nb-doped $SrTiO_3$ [66] Ag_2S [67], Cu_2S [68], or highly conducting regions on SiO_2 [69]. However, recent studies have shown that STM can also be applied to atomic switch experiments on ion conductors like $RbAg_4I_5$ [70] or even on materials considered as macroscopic insulators like Ta_2O_5 [71]. The key issue in this approach is using the switching time as a kinetic parameter that is independent of the electronic conductivity of the samples (in contrast to the current). Thus, one can increase the electronic conductivity of the ionic conducting films to a certain level (either by extrinsic doping or by thermal reduction), enabling the quantum mechanical tunneling and therefore STM, without significantly influencing the ionic processes.

As an example for such an atomic switch experiment, in which the electric field of a static STM tip is used to form a bridge of metal atoms between the sample surface and the tip, Fig. 28 shows the extraction of Ag atoms from a 150 nm thick $AgRb_4I_5$ film [70]. The number of atoms, which contact the STM tip, is given by number of quantum conductance steps $G_0 = 2e^2/h$ in the current-time diagram.

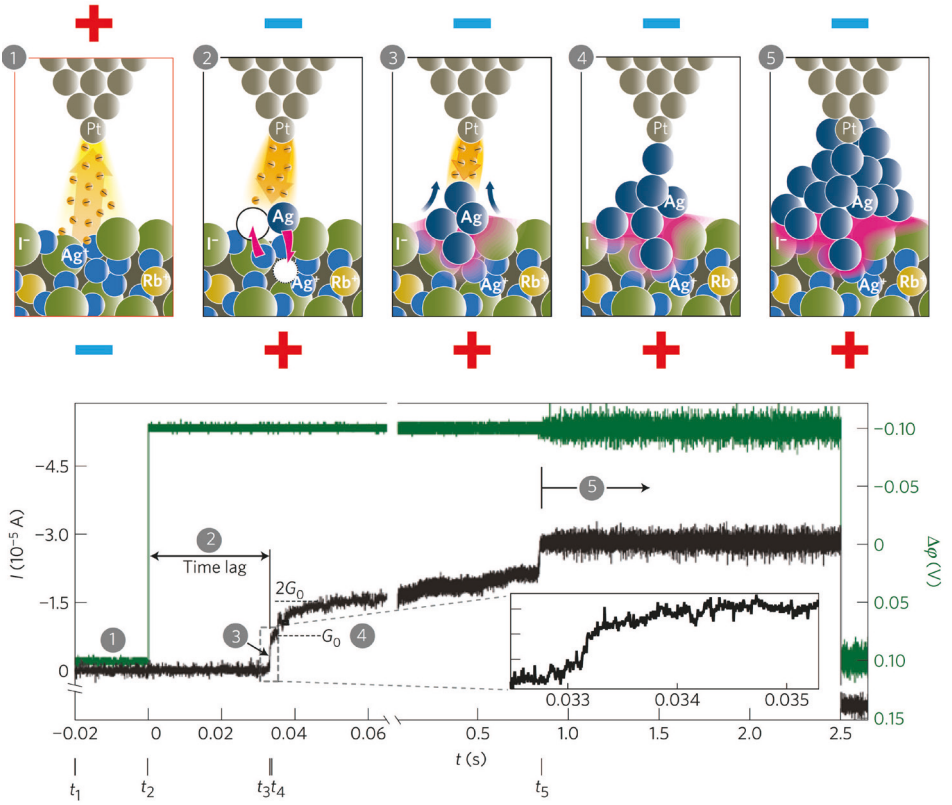


Fig. 28: *Current-time dependence at an applied voltage of −100 mV. t_i (i=1–5) denotes the starting time of the i^{th} process step. The number (1–5) notation relates the 5 regions defined in the current-time characteristics to the microscopic model (upper part of the graph) for the sequence of individual physicochemical processes during the switching. Only 4 Ag atoms in a chain are sufficient to short-circuit a gap of 1 nm (typical tip-sample distance). Accounting for dispersion of the tunnelling current beam area up to 20 Ag atoms can be calculated to constitute a cluster of a conical or tetrahedral form. The inset shows the first quantum step at G_0. The sharpness of the current increase is determined by the rate of the filament growth [70].*

5 Summary

The application examples presented here, illustrate only a very small fraction of the potential of the group of scanning probe microscopes. Nevertheless, the examples show that the scanning probe microscopes developed within a rather short period of time to indispensable tools in surface sciences, physics, chemistry, biology, materials development, and even technological developments in industry. The wide applicability is primarily due to the simple principle of scanning a probe tip over a surface and obtaining an image based on a probe tip – sample interaction. Depending on the type of interaction, surfaces can even routinely imaged with atomic resolution, providing a deep insight into the physical processes at surfaces. These advantages were to date not achievable using other techniques, which make the scanning probe techniques so unique.

References

[1] G. Binnig, H. Rohrer, Ch. Gerber, and E. Weibel, Appl. Phys. Lett. **40**, 178 (1982), Phys. Rev. Lett. **49**, 57 (1982).

[2] Ph. Ebert, K. Urban, and M.G. Lagally, Phys. Rev. Lett. **72**, 840 (1994).

[3] R. Wiesendanger, Rev. Mod. Phys. **81**, 1495 (2009).

[4] G. Binnig, C. F. Quate, and Ch. Gerber, Phys. Rev. Lett. **56**, 930 (1986).

[5] R. Wiesendanger, H.-J. Güntherodt, G. Güntherodt, R. J. Gambino, and R. Ruf, Phys. Rev. Lett. **65**, 247 (1990).

[6] D. Pohl, W. Denk, and M. Lanz, Appl. Phys. Lett. **44**, 654 (1984)

[7] G. Gamov, Z. Phys. **51**, 204 (1928).

[8] E. U. Condon and R. W. Gurney, Nature **122**, 439 (1928)

[9] L. Nordheim, Z. Phys. **46**, 833 (1927).

[10] P. Eckle, M. Smolarski, P. Schlup, J. Biegert, A. Staudte, M. Schöffler, H. G. Muller, R. Dörner, and U. Keller, Nature Physics **4**, 565 (2008).

[11] J. Bardeen, Phys Rev. Lett. **6,** 57 (1961).

[12] J. Tersoff and D.R. Hamann, Phys. Rev. **B 31**,805 (1985).

[13] J. Tersoff and D.R. Hamann, Phys. Rev. Lett. **50**, 1998 (1983).

[14] N. D. Jäger, E. R. Weber, K. Urban, and Ph. Ebert, Phys. Rev. B **67**, 165327 (2003).

[15] J.A. Stroscio, R.M. Feenstra, D.M. Newns, and A.P. Fein,J.Vac. Sci. Technol. A **6**, 499 (1988).

[16] L. Ivanova, S. Borisova, H. Eisele, M. Dähne, A. Laubsch, and Ph. Ebert, Appl. Phys. Lett. **93**, 192110 (2008).

[17] Ph. Ebert, L. Ivanova, and H. Eisele, Phys. Rev. B **80**, 085316 (2009).

[18] S. Landrock, Y. Jiang, K.H. Wu, E.G. Wang, K. Urban, and Ph. Ebert, Appl. Phys. Lett. **95**, 072107 (2009).

[19] Ph. Ebert, B. Engels, P. Richard, K. Schroeder, S. Blügel, C. Domke, M. Heinrich, and K. Urban, Phys. Rev. Lett. **77**, 2997 (1996).

[20] R.M. Feenstra, J.A. Stroscio, J. Tersoff, and A.P. Fein, Phys. Rev. Lett. **58**, 1192 (1987).

[21] Ph. Ebert, G. Cox, U. Poppe, and K. Urban, Surfac. Sci. **271**, 587 (1992).

[22] A. Selloni, P. Carnevali, E. Tosatti, and D.C. Chen, Phys. Rev. B **31**, 2602 (1985), idem **34**, 7406 (1986).

[23] J.A. Stroscio, R.M. Feenstra, and A.P. Fein, Phys. Rev. Lett. **57**, 2579 (1986).

[24] C.J. Chen, J.Vac.Sci. Technol. A **6**, 319 (1988).

[25] N. D. Jäger, K. Urban, E. R. Weber, and Ph. Ebert, Phys. Rev. B **65**, 235302 (2002).

[26] N. D. Jäger, *Effect of individual dopant atoms on the electronic properties of GaAs investigated by scanning tunneling microscopy and spectroscopy*, Thesis, RWTH-Aachen (2003).

[27] K. Besocke, Surf. Sci. **181**, 145 (1987).

[28] Y. Kuk and P.J. Silverman, Rev. Sci. Instrum. **60**, 165 (1989).

[29] H.-J. Güntherodt and R. Wiesendanger, *Scanning Tunneling Microscopy*, Vol. 1 and R. Wiesendanger and H.-J. Güntherodt, *Scanning Tunneling Microscopy*, Vols. **2** and **3**, Springer Series in Surface Science Vols. **20**, **28**, and **29**, Ed. R. Gomer, Springer, Berlin, 1992, 1993.

[30] G. Cox, *Untersuchung von Grenzflächen und Gitterbaufehlern in GaAs mit Hilfe der Rastertunnelmikroskopie*, Thesis, RWTH Aachen, published as Forschungszentrum Jülich GmbH Bericht 2382 (1990).

[31] S. Chiang (Ed.), Special Issue of Chemical Reviews **97** (4), June 1997.

[32] Ph. Ebert, Surf. Sci. Rep. **33**, 121 (1999); Current Opinion in Solid State and Materials Science **5**, 211 (2001).

[33] Ph. Ebert, M. Heinrich, M. Simon, C. Domke, K. Urban, C.K. Shih, M.B. Webb, and M.G. Lagally, Phys. Rev. B **53**, 4580 (1996).

[34] Ph. Ebert, M. Heinrich, M. Simon, K. Urban, and M.G. Lagally, Phys. Rev. B **51**, 9696 (1995).

[35] R. B. Dingle, Phil. Mag. **46**, 831 (1955).

[36] N. D. Jäger, K. Urban, E. R. Weber, and Ph. Ebert, Appl. Phys. Lett. 82, 2700 (2003).

[37] R. M. Feenstra *et al.,* in: *Semiconductor Interfaces at the Sub-Nanometer Scale*, Eds. H. W. M. Salemink and M. D. Pashley (Kluwer Academic, Dordrecht, 1993), p. 127; R. M. Feenstra *et al.,* Appl. Phys. Lett. **61**, 795 (1992).

[38] N. D. Jäger, M. Marso, M. Salmeron, E. R. Weber, K. Urban, and Ph. Ebert, Phys. Rev. B **67**, 165307 (2003).

[39] M. Bode, Rep. Prog. Phys. **66**, 523 (2003).

[40] S. F. Alverado and P. Renaud, Phys. Rev. Lett. **68**, 1387 (1992).

[41] D. T. Pierce, Phys. Scripta **38**, 291 (1988).

[42] D. Wortmann, S. Heinze, Ph. Kurz, G. Bihlmeyer, and S. Blügel, Phys. Rev. Lett. **86**, 4132 (2001).

[43] S. Blügel, D: Pescia, and P. H. Dederichs, Phys. Rev. B **39**, 1392 (1989).

[44] R. Ravlić, M. Bode, A. Kubetzka, and R. Wiesendanger, Phys. Rev. B **67**, 174411 (2003).

[45] M. Kleiber, M. Bode, R. Ravlić, and R. Wiesendanger, Phys. Rev. Lett. **85**, 4606 (2000).

[46] J. A. Stroscio, D. T. Pierce, A. Davies, R. J. Celotta, and M. Weinert, Phys. Rev. Lett. **75**, 2960 (1995).

[47] W. Wulfhekel and J. Kirschner, Appl. Phys. Lett. **75**, 1944 (1999).

[48] S. Heinze, M. Bode, A. Kubetzka, O. Pietzsch, X. Nie, S. Blügel, and R. Wiesendanger, Science **288**, 1805 (2000).

[49] Ph. Ebert, Rastersondenmikroskopie in: M. von Ardenne, G. Musiol, and U. Klemradt, Effekte der Physik und ihre Anwendungen. p. 215 (Verlag Harri Deutsch, 2005).

[50] J.N. Israelachvili, *Intermolecular and Surface Forces.* (Academic Press, 1985).

[51] D. Sarid, *Scanning Force Microscopy* (Oxford University Press, 1994).

[52] O. Wolter, Th. Bayer, and J. Greschner, J. Vac. Sci. Technol. A *9*, 1353 (1991).

[53] C. J. Chen, *Introduction to Scanning Tunneling Microscopy* (Oxford University Press, 1993).

[54] F. J. Giessibl, Rev. Mod. Phys. **75**, 949 (2003).

[55] A. Pfau and W. Schrepp, Physikalische Blätter **55**(6), 31 (1999).

[56] H. Krupp, Adv. Colloid Interface Sci. **1**, 111 (1967); R. H. French, J. Am. Ceram. Soc. **83**, 2117 (2000).

[57] R.-H. Robrock, *Rasterkraftmikroskopie*, IFF-Bulletin **37** (1990).

[58] M. Schneider, *Untersuchung mikromagnetischer Strukturen in einkristallinen Eisenschichten mit einem kombiniertem Kerr-Kraftmikroskop*, Thesis, Univ. of Köln, published as Berichte des Forschungszentrums Jülich 3059 (1995).

[59] E. Bonaccurso, F. Schönfeld, and H.-J. Butt, Phys. Rev. B **74**, 085413 (2006).

[60] F. M. Serry, K. Kjoller, J. T. Thorton, R. J. Tench, and D. Cook, Application Note 27, Rev A1, 6/1/04, Veeco Instruments Inc. (2004).

[61] F. Nardi, D. Deleruyelle, S. Spiga, C. Muller, B. Bouteille, D. Ielmini, J. Appl. Phys. **112**, 064310 (2012).

[62] C. Xu, S. Wicklein, A. Sambri, S. Amoruso, M. Moors, R. Dittmann, J. Phys. D: Appl. Phys. **47**, 034009 (2014).

[63] R. Münstermann, J.J. Yang, J.P. Strachan, G. Medeiros-Ribeiro, R. Dittmann, R. Waser, Phys. Status Solidi RRL **4**, 16 (2010).

[64] R. Münstermann, T. Menke, R. Dittmann, R. Waser, Adv. Mater. **22**, 4819 (2010).

[65] U. Celano, L. Goux, A. Belmonte, K. Opsomer, A. Franquet, A. Schulze, C. Detavernier, O. Richard, H. Bender, M. Jurczak, W. Vandervorst, Nano Lett. **14**, 2401 (2014).

[66] Y.L. Chen, J. Wang, C.M. Xiong, R.F. Dou, J.Y. Yang, J.C. Nie, J. Appl. Phys. **112**, 023703 (2012).

[67] A. Nayak, T. Tamura, T. Tsuruoka, K. Terabe, S. Hosaka, T. Hasegawa, M. Aono, J. Phys. Chem. Lett. **1**, 604 (2010).

[68] A. Nayak, T. Tsuruoka, K. Terabe, T. Hasegawa, M. Aono, Nat. Nanotechnol. **22**, 235201 (2011).

[69] A. Mehonic, S. Cueff, M. Wojdak, S. Hudziak, C. Labb, R. Rizk, A.J. Kenyon, Nat. Nanotechnol. **23**, 455201 (2012).

[70] I. Valov, I. Sapezanskaia, A. Nayak, T. Tsuruoka, T. Bredow, T. Hasegawa, G. Staikov, M. Aono, R. Waser, Nat. Mater. **11**, 530 (2012).

[71] A. Wedig, M. Luebben, D.-Y. Cho, M. Moors, K. Skaja, V. Rana, T. Hasegawa, K. Adepalli, B. Yildiz, R. Waser, I. Valov, Nat. Nanotechnol. (2015), doi:10.1038/nnano.2015.221.

D1 Magnetic Random Access Memory

Ioan Lucian Prejbeanu
SPINTEC, UMR 8191 CEA-CNRS-UGA-GINP, Grenoble, France

Contents

1 Introduction **2**
1.1 Status of emerging nonvolatile MRAM market 2
1.2 Current and Future Challenges for MRAMs 2

2 Field-written MRAM (FIMS-MRAM) **4**
2.1 Stoner-Wohlfarth MRAM 4
2.2 Toggle MRAM 6
2.3 Limitation in down-size scalability 8

3 Thermally-Assisted MRAM (TAS-MRAM) **8**
3.1 In-plane TAS-MRAM 8
3.2 TAS-MRAM with soft reference: Magnetic logic unit (MLU) 10

4 Spin-torque-transfer MRAM (STT-MRAM) **11**
4.1 Principle of STT writing 11
4.2 Considerations of breakdown, write, read voltage distributions 13
4.3 In-plane STT-MRAM 13
4.4 Perpendicular STT-MRAM 14

5 Thermally Assisted STT-MRAM **15**
5.1 In-plane TAS-STT-MRAM 15
5.2 Out-of-plane STT-TAS-MRAM 16

6 Conclusions **17**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

1.1 Status of emerging nonvolatile MRAM market

Demand for on-chip memories has been recently increasing due the growth in demand for data storage and the increasing gap between processor and off-chip memory speeds. One of the best solutions to limit power consumption and to fill the memory gap is the modification of the memory hierarchy by the integration of non-volatility at different levels (storage class memories, DRAM main working memory, SRAM cache memory), which would minimize static power as well as paving the way towards normally-off / instant-on computing (logic-in-memory architectures). Besides computers, today's portable electronics have become intensively computational devices as the user interface has migrated to a fully multimedia experience. To provide the performance required for these applications, the actual portable electronics designer uses multiple types of memories: a medium-speed random access memory for continuously changing data, a high-speed memory for caching instructions to the CPU and a slower, nonvolatile memory (NVM) for long-term information storage when the power is removed. Combining all of these memory types into a single memory has been a long-standing goal of the semiconductor industry, as computing devices would become much simpler and smaller, more reliable, faster and less energy consuming. As a result, advanced NVM chips are expected to see phenomenal growth in the forthcoming years. MRAM is one of a number of new technologies aiming to become a "universal" memory device applicable to a wide variety of functions. MRAMs are expected to combine nonvolatility, high speed, moderate power consumption, infinite endurance, and radiation hardness, all at moderate cost and be easy to embed in devices.

Since its inception in the late 1990s, MRAM have however not yet reached large volume applications, with only Toggle switching-based standalone products currently available from Everspin, at the 180 nm technology node [1]. The more recent advent of STT-MRAM, however, has shed a new light on MRAM with the promises of much improved performances and greater scalability to very advanced technology node. Indeed, in 2010, the International Technology Roadmap for Semiconductors (ITRS), Emerging Research Devices and Emerging Research Materials Working Groups *"identified spin transfer torque MRAM and redox RRAM as emerging memory technologies recommended for accelerated research and development leading to scaling and commercialization of nonvolatile RAM to and beyond the 16 nm generation"* [2]. Currently there is an intense research and development effort in microelectronics on these two technologies, one based on spintronic phenomena, the other based on migration of vacancies or ions in an insulating matrix driven by oxidation-reduction potentials. Both technologies could be used for standalone or embedded applications. As a consequence, MRAM is now viewed as a credible replacement for existing technologies for applications where the combination of nonvolatility, speed, and endurance is key.

1.2 Current and Future Challenges for MRAMs

The elementary cell of all MRAM architectures is a magnetic tunnel junction (MTJ) consisting of two ferromagnetic layers separated by a thin insulating barrier. Between 1996 and 2004, most research and development focused on MRAM written by field (top line in Fig. 1). Until the discovery of STT switching and its gradual implementation in MTJ after 2006, the only known way to manipulate the magnetization of a magnetic nanostructure (here the MTJ storage

layer) was indeed with use of a magnetic field. The magnetic field is created by pulses of current flowing in conducting lines located below and above the MTJ. These approaches were used in the first MRAM products (1, 4, 8, and 16 Mbit MRAM chips) by Freescale Semiconductor and its spin-off Everspin Technologies in 2006. An extension of the initial field written MRAM (Fig. 1a) is the Thermally Assisted MRAM (TAS-MRAM) [3] (Fig. 1b), mainly developed by Crocus Technology. In the latter, the write selectivity is achieved by a combination of temporary heating of the selected cell produced by the tunneling current flowing through the cell and a single pulse of magnetic field. The power consumption to write these memory elements is significantly reduced compared to conventional field-written MRAM thanks to the possibility of using lower magnetic fields and of sharing each field pulse among several cells so as to write several bits at once. Field-written technology is robust and is already used in a variety of applications where reliability, endurance, and resistance to radiation are important features, such as in automotive and space applications. However, the down-size scalability provided by field-writing in conventional technology is limited to MTJ dimensions on the order of 60nm×120nm due to electromigration in the conducting lines used to generate the field. In addition, in field-writing, the write field extends all along the conducting line where it is produced and decreases relatively gradually in space, inversely proportional to the distance to this line. As a result, unselected bits adjacent to selected bits may sense a significant fraction of the write field, which may yield accidental switching of these unselected bits. Besides, TAS-MRAM with a soft reference allows introducing new functionalities, particularly promising for security and routers applications. However, the downsize scalability in conventional field-writing technology is limited to about 60nm, due to electromigration issues in the field lines. The interest in using STT as a new write approach in MRAM has increased, motivated by the fact that STT-writing (Fig. 1c) offers a much better down-size scalability than field-writing as the critical current for writing decreases proportionally to the cell area down to a minimum value set by the retention (~15μA). Furthermore, STT provides very good write selectivity since the STT current flows through only the selected cells. The greatest interest is now focused on out-of-plane magnetized STT-MRAM, taking advantage of the perpendicular magnetic anisotropy which exists at the CoFeB/MgO interface (Fig. 1d) [4]. Perpendicular STT-MRAMs require significantly less write current than their in-plane counterparts for a given value of memory retention and provide a better stability of the written information. Optimized perpendicular STT-MRAM stacks will likely comprise two tunnel barriers with antiparallel polarizing layers to maximize anisotropy and STT efficiency (Fig. 1e). The thermal assistance can also be combined with STT to circumvent a classical dilemma in data storage between the memory writteability and its retention [5]. Recently it has been shown that assistance by an electric field may reduce the STT writing critical currents in MTJs [6]. This has been demonstrated in magnetic stacks with perpendicular magnetic anisotropy but the effect is quite weak when using metallic layers due to the electric field screening over the very short Fermi length in metals. An electrically reduced magnetic anisotropy leads to lower energy barrier that is easier to overcome for changing magnetization direction. In principle, voltage control spintronic devices could have much lower power consumption than their current-controlled counterparts provided they can operate at sufficiently low voltage (below 1V). Multiferroïc or ferromagnetic semiconductor materials could provide more efficient voltage controlled magnetic properties. 3-terminal MRAM cells written by domain wall propagation (Fig. 1f) or SOT (Fig. 1g) were also recently proposed [7] to separate write and read current paths. This can ease the design of non-volatile logic circuits and increase the reliability of the memory. SOT-MRAM offers the same non-volatility and compliance with technological nodes below 22nm, with the addition of lower power consumption, cache-compatible high speed and improved endurance. The drawback is the increased cell size.

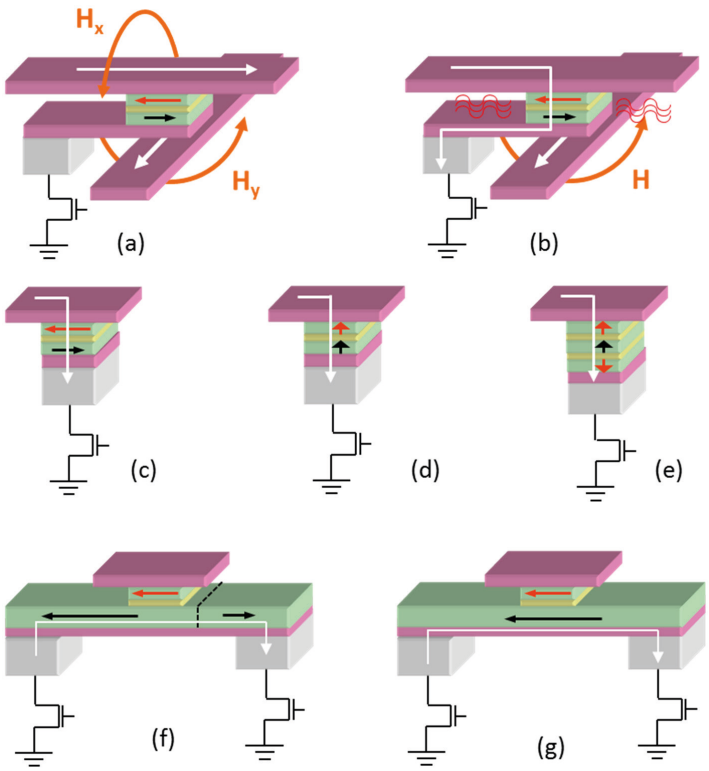


Fig. 1: *Various MRAM technologies: Toggle (a), Thermally Assisted MRAM (b), in-plane (c) and (d) out-of-plane magnetized STT-MRAM, Perpencular STT-MRAM with double barrier(e), 3-terminal devices based on domain wall propagation (f) and SOT (g) – reprinted from R.L. Stamps et al, J. Phys. D – Appl. Phys. 47(33) 333001 (2014)*

2 Field-written MRAM (FIMS-MRAM)

Two categories of field-induced magnetic switching MRAM (FIMS-MRAM) are described here below: Stoner-Wohlfarth MRAM and the "toggle" MRAM.

2.1 Stoner-Wohlfarth MRAM

The Stoner-Wohlfarth MRAM (SW-MRAM) was the first developed category of MTJ-based MRAM. The research and development in this area has been very useful in starting the development of hybrid CMOS/MTJ technology but it did not yield a product because of write selectivity problems and poor down-size scalability. SW-MRAM consists of an array of MTJs

in which each individual MTJ is connected in series with a selection transistor (Fig. 2). The MTJs are sandwiched between two sets of orthogonal conducting lines (bit lines and word lines) aimed at creating local magnetic fields on the MTJs storage layer when current flows are sent along them. To write at a particular addressed cell, two simultaneous pulses of current are sent in the bit line and world line which cross each other at the addressed MTJ cell. These currents must be adjusted so that the resulting field at the addressed cell is locally large enough to switch its storage layer magnetization in the desired direction while not switching the storage layer magnetization in the other memory points located further along the same bit line or the same world line. Indeed these other memory cells also feel the field created by the current pulse but not the two perpendicular fields simultaneously. They are called half-selected bits.

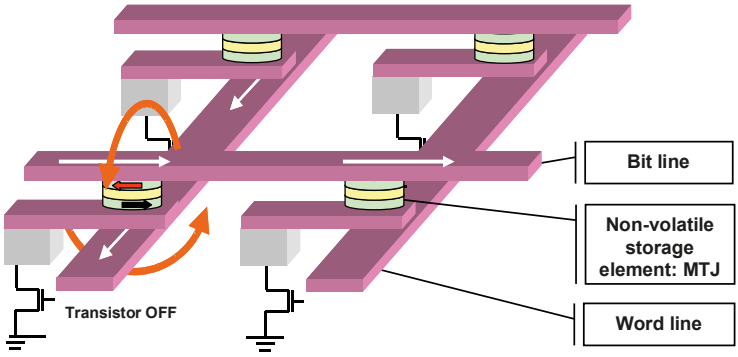


Fig. 2: *Schematic representation of a SW-MRAM array. The MTJ, patterned as elliptical cylinders are in-plane magnetized with one layer of fixed magnetization pinned along the ellipse long axis ("reference layer," black arrow) and one layer of switchable magnetization having two stable states along the ellipse long axis ("storage layer," red arrow). To address the memory element located at the front left of the array, two pulses of current (represented by white arrows) are simultaneously sent in the bit line and word line which cross each other at the addressed memory point. These pulses generate two perpendicular magnetic fields (orange arrows) which add as two vectors at the addressed memory point.*

In these SW-MRAMs, the write selectivity is thus based on the combination of two orthogonal magnetic fields, one along the easy axis of magnetization, the other along the hard axis. The write principle is based on the so-called Stoner-Wohlfarth switching astroid which provides a quantitative criterion for magnetization switching in a magnetic nanostructure with uniaxial anisotropy. A magnetic nanostructure of magnetization M_s has a uniaxial anisotropy described by the anisotropy energy per unit volume K_u. This nanostructure is assumed to be sufficiently small so that its magnetization remains homogeneous and therefore can be described in the macrospin approximation. We assume that the magnetization is initially oriented in one direction along the easy axis of magnetization and that we want to switch it in the opposite direction. This is achieved by applying simultaneously a field along the easy axis of magnetization (H_x) and another one in the orthogonal direction, i.e., along the hard axis of magnetization (H_y). The condition for switching is that the total field vector of the (H_x, H_y)

components must fall out of the Stoner-Wohlfarth astroid, which is defined by the relationship:

$$H_x^{2/3} + H_Y^{2/3} = \left(\frac{2K_u}{M_s} \right)^{2/3} \tag{1}$$

This relationship sets a lower limit to the amplitude of the write field. In order to avoid half-selected bits to switch, the easy axis field must be lower than the anisotropy field H_k given by

$$H_K = \left(\frac{2K_u}{M_s} \right),$$

otherwise this field alone would switch all bits located along the corresponding word line. Thus, taking into account the lower limit set by the SW astroid and upper limit set by the anisotropy field, this defines the ideal operating window for SW-MRAM. However, in practical devices, several factors actually restrain the size of this operating window:

i. The SW astroid in elliptic MTJ of typical dimensions 100nm×200nm is often distorted due to micromagnetic configurations of the magnetization.

ii. The switching criterion given by the SW astroid is actually valid only at 0 K. For devices operating at ambient temperature, thermal activation can significantly assist the magnetic switching so that the operating window must be pushed further away from the theroretical astroid.

iii. Due to variability in the patterning process, cell-to-cell distributions in anisotropy field lead to cell-to-cell distribution in the shape and size of the SW astroid.

As a result, it was very difficult to find any write operating window in SW-MRAM chips, even those of moderate capacity (e.g., 1 Mbit). Fortunately, a solution to this problem called "toggle writing" and described in the next paragraph was found.

2.2 Toggle MRAM

Toggle MRAM are also written with magnetic fields but the structure of the MTJ storage layer and the synchronization of the two orthogonal pulses of magnetic field differ from SW-MRAM. The magnetization in the MTJ ferromagnetic layers is still in-plane and the cells are patterned in elliptical shape, providing uniaxial shape anisotropy with easy axis along the long axis of the ellipse. In toggle MRAM, the ellipses are oriented at 45° with respect to the bit lines and word lines to optimize the write process. The storage layer consists of a compensated synthetic antiferromagnet, i.e., two ferromagnetic layers of same magnetic moment antiferromagnetically coupled through a thin Ru spacer layer. At rest, the magnetic moments of these two layers lie antiparallel along their easy axis of magnetization. When a moderate field is applied to such a structure, a magnetic transition takes place at a field called "spin-flop field" between a configuration at low fields, wherein the magnetic moments of the two layers lie antiparallel along their common anisotropy axis, and a configuration above the spin-flop field, where they lie symmetrically with respect to the field direction in a scissor configuration. As a result, at low fields, the system has no net magnetic moment whereas above the spin-flop field, it acquires a net magnetic moment. The spin-flop field $H_{spin-flop}$ is given by the following expression [8]:

$$\mu_0 M_s H_{spin-flop} = 2\sqrt{K_{eff}\left(\frac{A}{t} + K_{eff}\right)} \quad (2)$$

where M_s is the saturation magnetization of the two ferromagnetic layers (assumed here to be identical), t is their thickness, K_{eff} is their effective anisotropy per unit volume mainly of shape origin and A is the amplitude of the interfacial antiferromagnetic coupling through the Ru spacer. Therefore, the spin-flop field can be adjusted by varying the cell aspect ratio which determines K_{eff} or the ferromagnetic layers thickness or the Ru thickness which determines the amplitude of the antiferromagnetic coupling. The write toggle operation then proceeds as represented in Fig. 3.

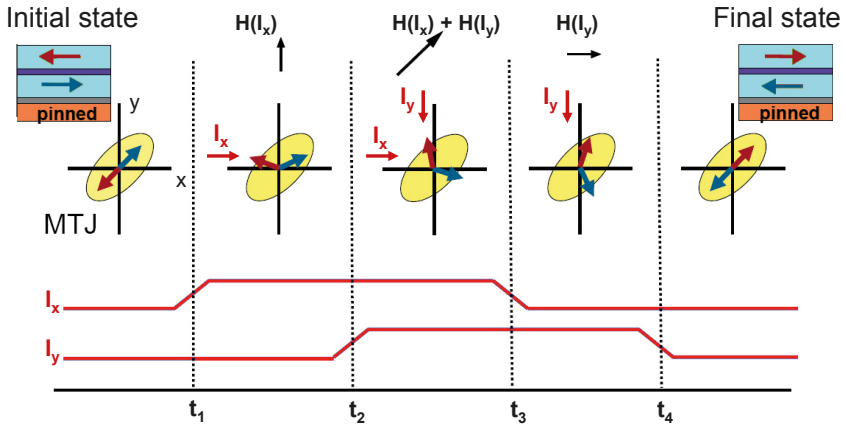


Fig. 3: *Toggle write operation sequence*

In the initial state, the two ferromagnetic layers are in antiparallel magnetic configuration along the long axis of the elliptical cell. At t_1, a current is sent in the x-line generating a magnetic field on the storage layer in the y-direction. The two magnetizations then scissor in the direction of this field and get in spin-flop configuration. The applied field is then gradually rotated by two steps of 45°. This is achieved by applying simultaneously a current along the x-line and y-line between t_2 and t_3 then only along the y-line between t_3 and t_4. During these steps, the spin-flop magnetic configuration rotates with the field. The y-current is then stopped. Since no more magnetic field is applied, the magnetization of the ferromagnetic magnetic layers then relax back to the antiparallel configuration along their easy axis of magnetization. In the final state, both layers have rotated by 180°. Toggle writing provides a much wider operation window than SW writing as the energy barrier to switch half-selected bits is many times larger than for switching fully selected bits [9]. Thanks to these improved write performances, Everspin succeeded in launching the first MRAM products on the market in 2006.

2.3 Limitation in down-size scalability

In field-written MRAM, down-size scaling is mainly limited by electromigration taking place in the field-generation word and bit lines. Indeed, in order to insure sufficient memory retention, the volume of the storage layer must fulfill the relationship $K_{eff}V>70k_BT$ (in which V is the volume of the storage layer and K_{eff} its effective anisotropy). As the feature size F decreases, the volume of the storage layer decreases as F^2 so that the effective anisotropy must be increased as F^{-2}. In toggle writing, the spin-flop field given by equation 6 roughly scales as $(K_{eff})^{1/2}$, i.e., as F^{-1}. Because the distance between the center of the field line and the storage layer does not vary significantly as the technology shrinks, the current required to generate the write field is proportional to the write field amplitude and therefore scales as F^{-1}. In terms of current density, if only the width of the field generating lines decreases while their height is approximately constant, the current density in these lines increases as F^{-2} when F itself decreases. Since the electromigration threshold in Cu is $\sim 10^7 A/cm^2$, this limits the MTJ width to dimensions above 100nm.

3 Thermally-Assisted MRAM (TAS-MRAM)

The magnetic field required to switch the magnetization is proportional to the anisotropy. Therefore, the larger the anisotropy, the larger the write field and therefore the larger the current required to create this magnetic field, meaning larger write power consumption. This general difficulty validates the interest in assisting the write, namely using a thermally assisted write approach. Indeed, in magnetic materials, it is known that generally it is easier to switch the magnetization of a magnetic element at elevated temperature than at low temperature. This originates from the fact that the magnetic anisotropy decreases with temperature so that the barrier height for switching decreases with temperature. Furthermore the higher thermal activation itself may help the magnetization to switch above the barrier. The concept of thermally assisted writing thus consists in storing the information at a standby temperature at which the anisotropy and therefore the thermal stability factor are very large and then temporarily increasing the temperature of the magnetic element during each write event to reduce the barrier height and ease the switching of the magnetization. A thermally assisted switching MRAM concept (TAS-MRAM) was proposed to improve the thermal stability, write selectivity, and power consumption of MRAM cells. [3]. In MRAM, the heating of the storage layer can be produced in a simple way by taking advantage of the Joule dissipation around the tunnel barrier. Actually the heating in MTJ is not a simple Joule heating as in metallic systems because we are here dealing with tunneling instead of ohmic transport. Rather, the heating in MTJ is due to the inelastic relaxation of tunneling hot electrons which takes place when the tunneling electrons penetrate in the receiving electrode after their ballistic tunneling across the tunnel barrier.

3.1 In-plane TAS-MRAM

In this scheme, a conventional MRAM stack is modified by replacing the simple ferromagnetic storage layer by an exchange biased storage layer, i.e., a ferromagnetic storage layer exchange coupled to an antiferromagnetic layer [3]. The write procedure requires heating above the storage layer blocking temperature and cooling in the presence of a magnetic field. The blocking temperature in a ferromagnetic/antiferromagnetic bilayer is the temperature at which

the loop shift induced by the exchange coupling across the interface between these two layers vanishes. The reference and the storage layer must be exchange biased with antiferromagnets having sufficiently different blocking temperatures: typically PtMn with blocking temperature of 350°C are used in the reference layer whereas the storage layer antiferromagnet is chosen with a blocking temperature in the range 180°C-250°C depending on the requirements on the device operating temperature range. The main advantages of the TAS-MRAM writing scheme are (1) the use of a single field selection line instead of two for toggle writing, (2) an important reduction in write power consumption and (3) a largely improved thermal stability. One way to increase the heating efficiency and reduce the power density is to insert low thermal conductivity materials at both ends of the magnetic tunnel junction stack, serving as thermal barriers between the junction and the electrical leads. This confines the heat to the junction volume, preventing lead heating and possible thermal crosstalk [10]. Typical heating times in TAS-MRAM are in the range 3 to 10ns, primarily influenced by the heating power dissipated at the tunnel barrier. The cooling rate depends on the heat diffusion constant towards the top and bottom of the MTJ stack and on the specific heat of the MTJ and typically ranges between 10 to 20ns [11].

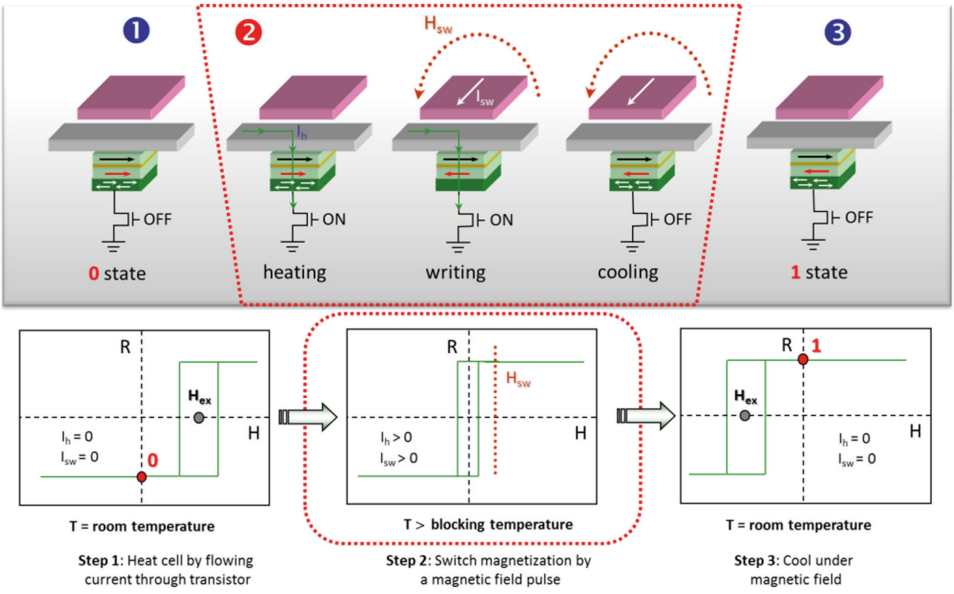


Fig. 4: *Writing steps in the Thermally Assisted MRAM architecture – Reprinted from Prejbeanu et al, J. Phys. D: Appl. Phys. 46 074002 (2013)*

Fig. 4 illustrates a TAS-MRAM bit write sequence example. The writing process starts from a given initial orientation of the magnetization of the exchange biased storage layer for instance representing a low resistance state "0". The corresponding storage layer loop is shifted around a negative field, in the hysteresis cycle before the heating pulse is applied. The reversal of the storage layer bias is achieved by heating the AF layer above its blocking temperature with a current pulse and applying simultaneously an external magnetic field H_{sw} larger than the coer-

cive field of the storage layer. The field is applied in a direction that favors the anti-parallel alignment of the storage and reference layers. The current pulse is terminated and the system is cooled in a magnetic field. The result is a reversal of the pinning orientation of the storage layer and a bit state change to a high resistance "1". As a result the storage layer loop is now shifted towards positive values. One unique feature of the TAS-MRAM approach is the protection against field erasure in standby. This means that, due to the exchange biasing of the storage layer, the P and AP resistances remain unchanged, even if the MTJ is subject to magnetic field perturbations. In standby, only one state of the storage layer is stable at zero field. This means that even if the TAS-MRAM chip is exposed to a perturbation field, this field may temporary switch the magnetic configuration of the memory but the latter will spontaneously return to its original state before perturbation once the perturbation disappears.

3.2 TAS-MRAM with soft reference: Magnetic logic unit (MLU)

In a second possible implementation of TAS-MRAM, a soft reference (SR) layer is made of a material with easily switchable magnetization (see Fig. 5). The storage layer is exchange biased by an adjacent antiferromagnetic layer as in standard TAS-MRAM. The writing of the storage layer is achieved similarly to thermally assisted MRAM devices.

The reading scheme is different with respect to the classical TAS-MRAM and consists in switching the free layer in a first predetermined direction (along the stable directions of storage layer magnetization) and then in the opposite direction [12]. This reading can be performed in a two-step quasi-static way or dynamically with an oscillation of the sense layer magnetization away from a 90° orientation, with the storage layer magnetization yielding an upward or downwards oscillation of the MTJ resistance. The quasi-static read is performed in two steps – Fig. 6:

1. The soft reference layer is first set in a predetermined initial direction by application of a first pulse of magnetic field (without heating pulse so as not to write the storage layer). Then the resistance of the MRAM cell is measured.
2. The soft reference magnetization is then switched to the opposite direction and the new resistance is measured.

In this approach, the read cycle is longer (~50ns) but the tolerance to process variation is greatly enhanced since each bit is self-referenced. Indeed, for the standard reading scheme, the two resistance state distributions have to be well separated in order to avoid read errors. This implies good dot size and shape control. In contrast, for the SR reading scheme, the dif-

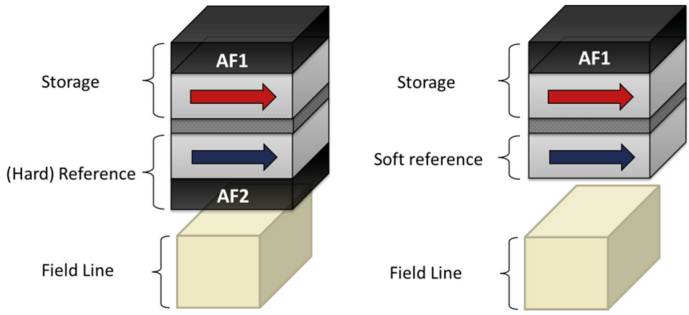


Fig. 5: *Schematic representation of a self-referenced MRAM.*

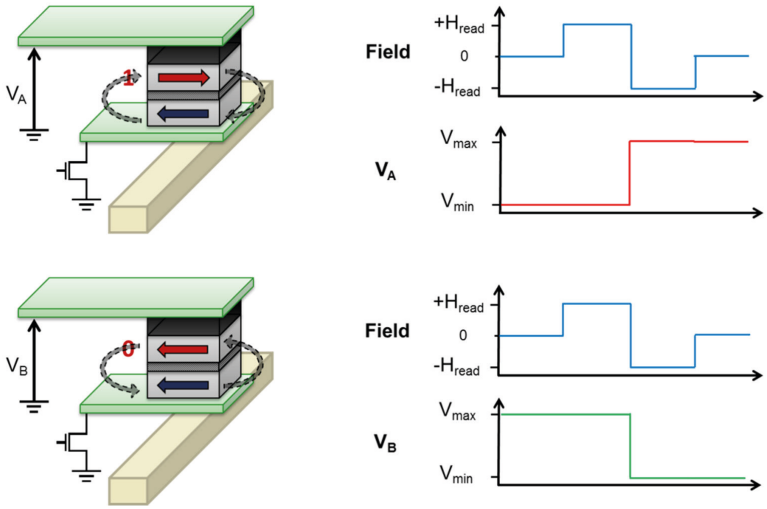


Fig. 6: *Reading procedure of a self-referenced MRAM*

ference of resistance between the two states is used to read the junction, and is thus not sensitive to dot size variation. This is particularly useful for small technological nodes where it becomes tricky to control accurately the resistance distributions.

Furthermore, the high blocking temperature antiferromagnetic material employed in usual MRAM reference layers is no longer required in SR-MRAM stacks leading to a large increase of the operating temperature range. Thanks to the use of only one antiferromagnet, one of these boundaries is removed and the programming range can extend to much higher temperature since there is no more risk to unpin the reference layer during write. Antiferromagnets with higher blocking temperature may also be used to pin the storage layer magnetization for high temperature applications. SR-MRAM cells, aside from their storage functionality, have also been identified as technological solutions for high-temperature, multi-dimensional field sensing applications [13], as well as power amplification [14].

4 Spin-torque-transfer MRAM (STT-MRAM)

4.1 Principle of STT writing

When a spin-polarized current flows through a magnetic nanostructure, the STT results from the interaction between the spin of the conduction electrons and those responsible for the nanostructure magnetization. This torque is exerted on the local magnetization and tends to switch it towards a direction parallel or antiparallel to that of the spin polarizing layer depending on the current direction. Taking into account the STT, the general equation governing the dynamics of magnetization of the magnetic nanostructure is written as:

$$\frac{d\hat{m}}{dt} = -\gamma\left(\hat{m}\times\vec{H}_{eff}\right) + \alpha\left(\hat{m}\times\frac{d\hat{m}}{dt}\right) + \frac{\gamma\hbar}{2\mu_0 M_s}\frac{1}{d}\frac{\eta J}{e}\hat{m}\times\left(\hat{m}\times\hat{p}\right) \qquad (3)$$

In this extension of the Landau – Lifshitz - Gilbert (LLG) equation, the first term describes a precessional motion of the magnetization around the local effective field $\vec{H}_{eff}$ which contains contributions from the applied field, the demagnetizing field, the anisotropy field, and the STT field-like term. This first term is conservative, while the second term is the Gilbert damping term which describes the magnetic dissipation in the system, i.e., the fact that in the absence of STT influence, the magnetization tends to gradually relax towards the local effective field. Depending on the current direction through the magnetic nanostructure, this term can absorb or dissipate energy, which means that it either behaves as a damping or antidamping term. If the current has the proper direction and the current density is large enough, the STT antidamping effect can exceed the natural Gilbert damping. Very peculiar magnetization dynamics effects then arise, such as STT-induced magnetization switching or magnetization steady-state oscillations [15]. STT offers a new way to manipulate the magnetization of magnetic nanostructures. STT-induced magnetization switching provides a new way to write the information in MRAM or logic devices. STT-induced steady-state magnetic oscillations allow the generation of RF voltage and thus new types of frequency-tunable RF oscillators.

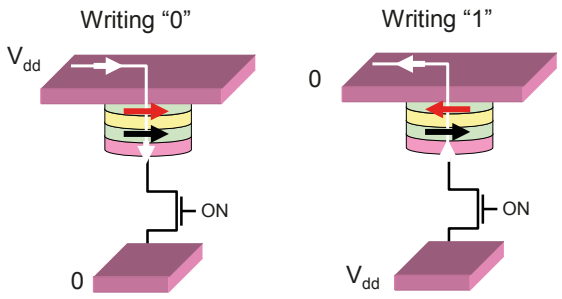


Fig. 7: *Principle of writing in STT-MRAM. Each STT-MRAM cell consists of an MTJ connected in series with a transistor. To write the parallel magnetic configuration, a current flow is sent through the MTJ from the storage layer (red arrow) to the pinned reference layer (black arrow) To write the antiparallel magnetic configuration, a current flow is sent through the MTJ from the reference layer (red arrow) to the storage layer.*

The writing is performed with bipolar pulses of current. The reading is performed at lower current to avoid write errors during read. Writing a “0” (i.e., a parallel configuration of the magnetization in the storage and pinned layers) can be achieved by sending a current pulse through the stack, the electrons flowing from the pinned layer to the storage layer. Writing a “1” can be achieved by sending a pulse of current of opposite polarity. STT indeed provides a powerful write scheme in MRAM for several reasons:

- In STT-MRAM there is no need to create pulses of magnetic field. Each cell is directly written by the current flowing through the stack. As a result the cells are much more compact than in field-written MRAM, as illustrated in Fig. 7.
- This approach clearly solves the write selectivity problem of the field written MRAM since the write current flows only through the addressed cell so there is no risk of writing an unselected cell.
- In STT writing, the condition for magnetization switching is set by a critical current density j_c. The magnetization of the storage layer switches if the current density of proper direction exceeds j_c. This provides very good down-size scalability since the total current required to write scales like the cell area down to very small dimensions where it becomes limited by the thermal stability factor.

However, at very small dimensions, there are still problems concerning the thermal stability of the information written in the cell.

4.2 Considerations of breakdown, write, read voltage distributions

In conventional STT-MRAM, the write and read current paths are the same. In order to avoid write disturbance during read, the read voltage must be chosen low enough compared to the critical write voltage. There are therefore 3 voltage cell-to-cell distributions in an MRAM chip which need to be well separated for proper functioning and reliability of the chip. These 3 distributions are:

i. **The breakdown voltage distribution.** The MTJ tunnel barrier is a thin dielectric oxide layer (MgO ~1nm thick). When exposed to an excessively large voltage, this barrier may experience dielectric breakdown.

ii. **The write voltage distribution**. At each write event (and to lesser extend read event), the tunnel barrier is exposed to an electrical stress which may cause electrical breakdown. To avoid breakdown failure, the highest write voltage in the distribution must be sufficiently low compared to the weakest MTJ in terms of breakdown. By adjusting the MTJs stack composition and their RA, one tries to get this write voltage distribution centered around 0.5V and be as narrow as possible. The distribution width mainly originates from fluctuations in the shape and particularly in edge defects associated with the patterning process.

iii. **The read voltage distribution.** The read voltage distribution originates from the variation in the resistance of the selection transistor which is connected in series with the MTJ. The read voltage across the MTJ is typically in the range 0.1 to 0.15 V and must be low enough compared to the write voltage in order to avoid any write disturbance during read caused by the STT from the read current. However, the lower the read voltage, the slower the read-out process. Therefore, a trade-off must be found.

4.3 In-plane STT-MRAM

The expression of the critical current for switching obtained in a macrospin model and at zero temperature is [16]:

$$J_{c0} = \frac{2e\alpha\mu_0 M_s t_F \left(H + H_k + \frac{M_s}{2} \right)}{\hbar\eta} \tag{4}$$

In this equation, H is the applied field along the easy axis, M_s and t_F the magnetization and the thickness of the storage layer, α is the damping constant, H_k the in-plane anisotropy field, η the spin transfer efficiency. When the injected current has the proper direction (see Fig. 7) and is larger than the critical current, the magnetization reverses. In this expression, the term $M_s/2$ is usually much larger than $(H+H_k)$ by one or two orders of magnitude. This dominant role of the demagnetizing field term $(\mu_0 M_s/2)$ comes from the fact that during the STT-induced switching of the magnetization of the in-plane magnetized layer, the magnetization has to precess out-of-plane which increases the demagnetizing energy. As a result, a good approximation of J_{c0} is given by

$$J_{c0} = \frac{e\alpha\mu_0 M_s^2 t_F}{\hbar\eta}.$$

It is also important to note that because of this dominance of the demagnetizing field term, the critical current for STT writing weakly depends on H_k, the in-plane anisotropy field which determines the thermal stability of the magnetization at rest.

4.4 Perpendicular STT-MRAM

The critical current for spin transfer reversal of the storage layer obtained from the LLG equation in the out-of-plane configuration is given by:

$$I_{c0} = \frac{2\cdot e}{\hbar}\cdot\frac{\alpha\cdot A\cdot t\cdot\mu_0 M_s}{\eta}\cdot H_{eff} \tag{5}$$

where A is the area of the magnetic element, e is the electron charge, $\hbar$ is the reduced Planck constant, μ_0 the vacuum permeability, α is the Gilbert damping coefficient, M_s and t are the saturation magnetization and thickness of the storage layer, η the spin transfer torque efficiency which depends on the relative orientation of the magnetizations (θ=0 or π) and on the polarization P, and H_{eff} the effective switching field. In magnetic junctions with out-of-plane magnetization, the effective field is given by:

$$H_{eff} = H_{K\perp} - M_s \tag{6}$$

where $H_{K\perp}$ is the perpendicular anisotropy field which pulls the magnetization out-of-plane ($H_{K\perp}>M_s$). In contrast to the in-plane magnetized case, both STT and thermal energy barriers are here identical. The critical switching current can thus be much smaller than for in-plane magnetized electrodes. By introducing the thermal stability factor

$$\Delta = \frac{KV}{k_B T} = \frac{\mu_0 M_s H_{eff} A t}{2k_B T} \tag{7}$$

where K is the anisotropy, $V=A\cdot t$ the volume of the storage layer, k_B is the Boltzmann's constant and T is the absolute temperature, relation (5) can be rewritten:

$$I_{c0} = \frac{4\cdot e}{\hbar}\cdot\frac{\alpha\cdot k_B T}{\eta}\cdot\Delta \tag{8}$$

This relation expresses that in the macrospin approximation, a direct proportionality exists between the thermal stability factor and the write critical current.

Perpendicular STT-MRAMs have many advantages compared to in-plane magnetized MTJs:

i. The switching current density is significantly reduced because the two terms present in the expression of the critical current density partially cancel [17].

ii. The thermal energy barrier is provided by this large effective perpendicular anisotropy instead of in-plane shape anisotropy. As a consequence, elongated cell shapes are no longer needed and the perpendicular MTJs can be patterned in circular shape. This facilitates manufacturability at smaller technology nodes and leads to smaller switching current for a given critical current density, resulting in smaller cell size.

iii. Finally, dipole field interaction between neighboring cells can also be reduced in high bit density layouts.

The perpendicular magnetic anisotropy (PMA) can have different origins, either bulk (in hcp CoCrPt, heavy rare earth-transition metal, or $L1_0$ FePt ordered alloys), or interfacial (in Pt/Co, Pd/Co, or Co/Ni multilayers). A large PMA can namely be induced at the interface between a ferromagnetic electrode and an oxide [18]. The main drawback of all these PMA-inducing materials resides in the fact that they generally induce an *fcc* crystallographic texture which is incompatible with the *bcc(001)* texture of the MgO crystalline barrier. The great breakthrough came in 2010 [19] when structures based on Ta/CoFeB electrodes were proposed, very similar to their in-plane counterparts. The problem of the bcc(100) texturation of the CoFeB electrode upon crystallization was thus solved.

A figure of merit has been proposed to characterize the efficacy of the STT [20], which is the ratio of the thermal stability factor to the critical switching current, Δ/I_{C0}. The figure of merit obtained with out-of-plane magnetized MTJ are expected to be much better than with in-plane magnetized MTJs. This is however true only if the Gilbert damping constant α, to which I_{c0} is proportional, can be maintained as low as in in-plane magnetized material, which is not very easy. Indeed, both Gilbert damping and magnetic anisotropy derive from the spin-orbit interactions which basically couple the electron spin to the lattice. As a result materials that exhibit large anisotropy (for instance FePt ordered alloys, CoPt multilayers and alloys) also exhibit large Gilbert damping constant α (in the range 0.05-0.2). A promising route to circumvent the problem of large damping in out-of-plane magnetized materials consists in using a storage layer where the perpendicular anisotropy does not arise from a bulk contribution but from a large interfacial perpendicular anisotropy which exists at the interface between the magnetic electrode and the tunnel barrier.

5 Thermally Assisted STT-MRAM

The same current flowing through the cell in STT-MRAM can be advantageously used both to heat up the cell and switch the storage layer magnetization by STT. Alternatively, thermal assistance can be used to extend the scalability of STT-MRAM both with in-plane and out-of-plane magnetized materials.

5.1 In-plane TAS-STT-MRAM

With in-plane magnetized materials, structures similar to the exchange biased storage layer stacks used for TAS-MRAM can be used – see Fig. 8. The resistance state was switched between the low resistance and high resistance states by applying current pulses of alternating polarity across the junction. Each pulse first creates a temperature increase above the antiferromagnet blocking temperature. With the ferromagnetic layer no longer pinned, the spin-polarized current simultaneously exerts a torque on the ferromagnetic storage layer reversing its magnetization direction depending on the current direction. The write voltage is then gradually decreased to zero so that the junction cools down while STT is still on. The antiferromagnet then freezes the new storage layer magnetization direction.

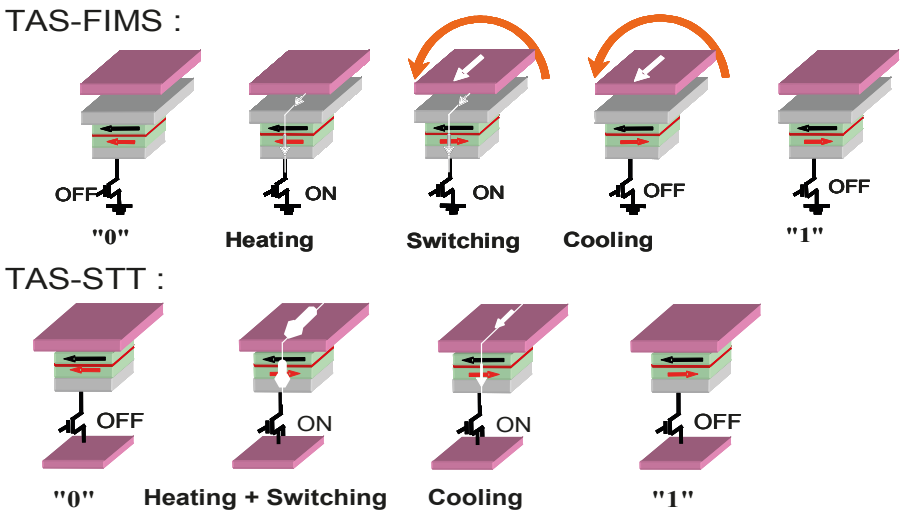


Fig. 8: *TAS write principle combined with field or STT. In TAS-STT-MRAM, the current flowing through the MTJ both heats the MTJ and exerts the magnetic torque which switches the magnetization. – reprinted from Dieny B et al, 2010 Int. J. Nanotechnol. 7 591*

5.2 Out-of-plane STT-TAS-MRAM

For technological nodes below 22nm, it becomes increasingly difficult to achieve sufficiently large effective anisotropy with in-plane magnetized as it is necessary to fulfil the 10 years data retention criterion. In order to solve this, perpendicular magnetic anisotropy is of great interest. However, current induced switching still present some issues which deteriorate the reliability of the devices. First, since the magnetization of the free layer is collinear to the spin polarization of the current in the initial state, a thermal fluctuation is required to initiate the reversal of the free layer by STT, leading to a stochastic switching. In addition, since the switching current is proportional to the effective anisotropy K^{eff} of the free layer, a dilemma has to be addressed: the decrease of the lateral dimensions of the device requires an increase of K^{eff} to maintain sufficient thermal stability. However, this increase in K^{eff} also yields higher write current and correlatively larger power consumption. This leads to a decrease of reliability since higher writing voltage reduces the existing margin to the breakdown voltage. One significant step forward towards high density memory cells has been to assist thermally the spin transfer switching [5]. This was achieved by a thermally induced reorientation of the free layer magnetic anisotropy from out-of-plane to in-plane. This allows the spin transfer torque efficiency to be maximized during write, reduces the STT switching current and suppresses stochastic variations in switching time. This thermal assistance scheme was demonstrated in a MRAM cell designing a magnetic electrode coupled to a Co/Pd multilayer having a strong temperature dependence of the perpendicular anisotropy – see Fig. 9.

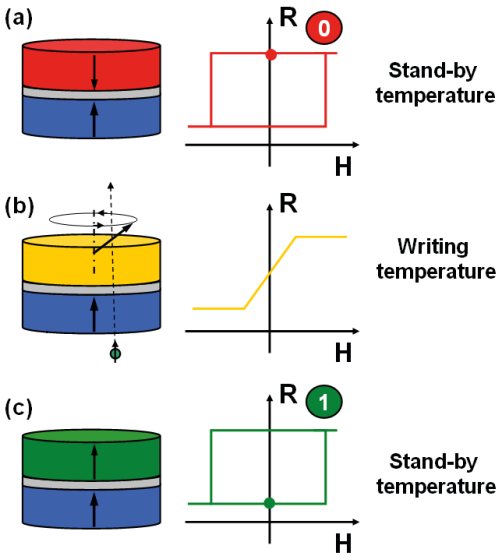


Fig. 9: *Principle of TIAR assisted switching: (a) the junction exhibits a large PMA at stand-by temperature. (b) A current is sent through the junction. The FL magnetization undergoes by heating a TIAR. The STT and pulls the FL magnetization upwards or downwards depending on the current direction. (c) The FL recovers its PMA during the cooling when the current is gradually decreased to zero. Reprinted from Bandiera S et al Appl. Phys. Lett. 201, 1 99 202507*

6 Conclusions

MRAMs are expected to combine nonvolatility, high speed, moderate power consumption, infinite endurance and radiation hardness, all at moderate cost and be easy to embed in devices. The potential benefits of spin-based memories are especially appealing when viewed in light of the exploding demand for on-chip memories. However, spin-based devices are still in their nascent stages, and in order to realize their potential, there is a need to strengthen the technology maturity and for advances in circuit designs and innovative architectures. There is still a need to strengthen the technology maturity and for advances in circuit designs and innovative architectures. The main issues remain associated with the cell to cell variability, TMR amplitude and temperature range. Variability is mainly caused by edge defects generated during patterning of the cells. MgO damages yield local changes in the barrier resistance, TMR and magnetic anisotropy i.e. cell retention. With the increasing number of actors now working on this technology, faster technological progresses can be expected in the near future. Also, implementing self-referenced reading scheme can lead to improved tolerance to process defects. Concerning out-of-plane STTRAM, progresses are needed in the composition of the stack to minimize the write current, maximize the TMR amplitude and improve the temperature operating range. Heusler and $X_{1-x}Mn_x$ (X = Cr, V, Ge, Ga…) alloys have also already demonstrated their potential for p-STTRAM (low M_s, large perpendicular anisotropy, low damping) [21] but none of the existing alloys combine all required properties yet. In particular, STTRAM has the potential of delivering high density and a scalable technology down to size ~20nm by using out-of-plane magnetized MTJ. Progress is also steadily being made in the composition of the stack to maximize the TMR amplitude, particularly in the perpendicular-MTJ configuration (now above 200%), and in improvements in the temperature operating range (to minimize the decrease of the PMA with operating temperature). The ultimate scala-

bility in STTRAM could be provided by combining thermally/voltage assisted switching and STT. SOT-MRAM can be viewed as a very interesting approach for non-volatile logic and MRAM of improved endurance. Voltage controlled spintronics devices may later yield devices of much reduced power consumption. As a matter of fact, there is lot of room for reducing the power consumption in MRAM technologies considering that the barrier height to insure a 10 year retention of a Gb chip is typically of $80k_BT \sim 4x10^{-4}$fJ whereas the energy presently required per STT write event is in the range 50fJ-1pJ.

References

[1] L. Savtchenko et al, US6545906 (2001).

[2] http://www.itrs.net/Links/2010ITRS/2010Update/ToPost/ERD_ERM_2010FINALReportMemory Assessment_ITRS.pdf

[3] B. Dieny and O. Redon, patent FR2832542, I.L. Prejbeanu et al, IEEE Trans. Magn 40 2625 (2004); I.L. Prejbeanu et al, J. Phys. D: Appl. Phys. 46 074002 (2013)

[4] M. Gajek et al, Appl. Phys. Lett. 100 132408 (2012)

[5] S. Bandiera et al, Appl. Phys. Lett. 99 202507 (2011)

[6] W.G. Wang et al, Nat. Mater., 11, 64 (2012)

[7] I.M Miron et al, Nature, 476 189 (2011)

[8] B. Dieny et al, J. Phys. Condens. Matter.2, 159 (1990)

[9] D.C. Worledge, Appl. Phys. Lett.84, 4559 (2004).

[10] S. Cardoso et al, J. Appl. Phys. 99(8) (2006)

[11] C. Papusoi et al, New Journal of Physics volume: 10 103006 (2208)

[12] N. Berger and J. P. Nozières, U.S. Patent 20110007561, issued January 13, 2011

[13] B. F. Cambou et al, September 19 2013. US Patent App. 13/787.585.

[14] I. L. Prejbeanu et al, March 26 2014. EP Patent App. EP20.120.290.316.

[15] W. H. Rippard et al, Phys. Rev. Lett. 92, 027201 (2004)

[16] Y. Huai, APPS Bulletin, vol. 18, No. 6 (2008)

[17] N. Masahiko et al, J. Appl. Phys. 103, 07A710 (2008).

[18] S. Monso et al, Appl. Phys. Lett. 80 (2002) 4157.

[19] S. Ikeda et al, Nature Materials 9 (2010) 721.

[20] T. Kishi et al, Int. Electron Devices Meeting Tech. Digest (2008)

[21] H.X. Liu et al, Appl. Phys. Lett. 101, 132418 (2012)

D2 Electrochemical Metallization Memories

Ilia Valov
Institute of Electronic Materials, PGI-7
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Materials **5**
2.1 Solid Electrolyte Materials 5
2.2 Electrode Materials 7
2.3 The nanobattery effect 10

3 Processes **15**
3.1 Switching Kinetics 15
3.2 Filament dynamics and quantum point contacts 16

4 Conclusions **20**

References **21**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

The modern nanoelectronics and information technology are facing the physical limitations of downscaling electronic elements. To address the demands on high information storage density, low power consumption, ultra-low write-erase-read times and non-volatility, a shift to new physical concept(s) is essential. The resistive switching memories (RRAM) are currently considered as major candidates to fulfil the requirements of the market being of interest for both interdisciplinary scientific community and industry. In more practical aspect RRAMs are emerging nanodevices with a great potential as a disruptive technology for the semiconductor industry as of a number of applications such as memory[1-3], logic[4-6] analog circuits[7], memristive operations, neuromorphic applications and computing[8-10]. These devices are non-volatile, with low power consumption, switching time down to sub-nanoseconds[11] and prospects for scalability approaching the atomic level[12]. Among other RRAM concepts those based on ion-conducting films and redox processes (ReRAM) show particular promise[1, 13-15]. The information in ReRAM is stored as different resistive states of nanoscale electrochemical cell consisting of two metal electrodes with a solid electrolyte in between. The transition between the low resistive ON-state (Boolean 1) and the high resistive OFF-state (Boolean 0) is provided either by formation and rupture of metallically conductive filament (filamentary type), or due to change of the interface properties (surface area type).

Filamentary type switching cells are in general prospective. One of the reasons is that the change in the absolute value of the resistance is high and, thus facilitating an easy practical differentiation between LRS (ON) and HRS (OFF), when downscaling the devices to nanometer scale. In Fig. 1 the three mostly favored ReRAM cells are presented, together with the corresponding current-voltage characteristics:

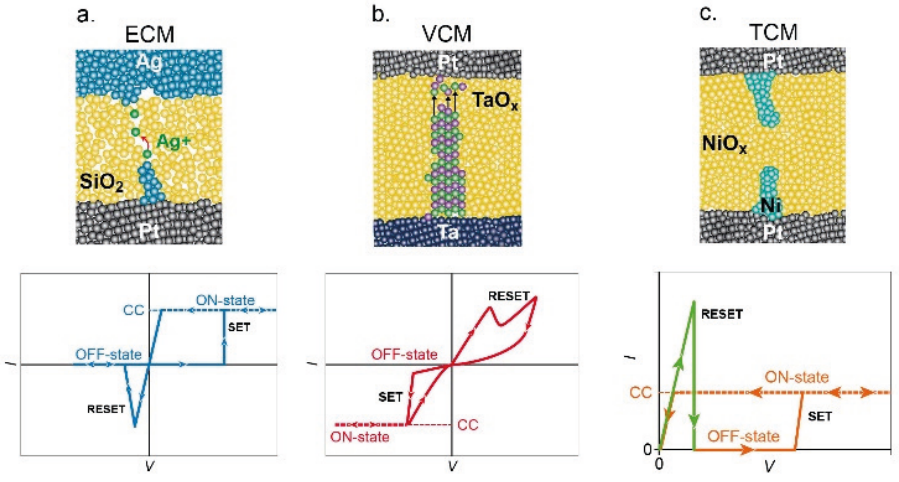


Fig. 1: *Type ReRAMs. a) Electrochemical Metallization Memory (ECM). b) Vacancy Change Memory (VCM) and c) ThermoChemical Memory (TCM). The I-V characteristics (below each cell type) are given referenced to the bottom electrode (grounded).*

A more precise classification of ReRAMs accounts for the switching mechanism and distinguishes electrochemical metallization memories, valence change memories and thermochemical memories abbreviated as ECM, VCM and TCM, respectively[16]. However, new studies have demonstrated that different switching mechanisms appeared related[17, 18].

The main distinctive feature of ECM and VCM is the bipolar type of switching i.e. the formation and dissolution of the filament occurs at different voltage polarity, whereas TCM memories are unipolar i.e. the filament formation/dissolution appears at the same polarity depending only on the particular voltage magnitude.

ECM cells have some particular advantages – they show bipolar characteristics and allow for the highest LRS/HRS ratio (up to 10^8). They are also often termed as conductive bridge RAM (CBRAM), programmable metallization cells (PMC), gapless type atomic switch etc.

The original term PMC now refers to the technology platform, which encompasses a wide range of applications beyond memory, including microelectromechanical systems [13], microfluidics [14], and optics [15]. During commercial development, the name CBRAM became popular and this is now typically used by the semiconductor industry exclusively for the memory variant [16, 17]. "Atomic switch" or "gap-type atomic switch" was introduced to distinguish sandwich type MIM cells from the cells where the switching event occurs in a vacuum gap between an electrode (e.g., a STM tip) and the solid electrolyte [18]. This terminology was further complemented by "gapless-type atomic switch" as alternative to ECM, CBRAM and PMC. In addition, some variants have been called "electrochemical memory" and the generic expressions "ionic memory" or "nano-ionic memory" have also been applied. The term ECM became used in the scientific literature when describing the underlying electrochemical processes [19].

The functional principle of a ECM cell is based on electrode redox reactions and nanoionic transport. In Fig. 2 the process of resistive switching in cell based on Cu^{z+}/Cu redox reaction and a Cu-ion conducting material is schematically shown.

The formation/dissolution of filaments in other types of systems is similar involving however different ions and electrode reactions.

One can also distinguish systems with and without a tunnel gap between one of the electrodes and the electrolyte, i.e. gap type atomic switch and gapless type[14]. As schematically shown in Fig. 3 in the gap type atomic switch the filament is formed not within the electrolyte but in the vacuum gap.

The atomic switch can be understood as a special case of ECM cell.

The most remarkable characteristic of the current state-of-the-art ReRAM cells are their dimensions scaling laterally below 10 nm side[21] and vertically down to some tens nanometer. These small dimensions predetermine physicochemical properties strongly deviating from the well-studied macroscopic systems and offer unique advantage of enabling using unconventional materials systems. For example SiO_2 at room temperature is a macroscopic insulator but reducing the thickness of the material down between 10 nm and 30 nm we observe Ag^+ or Cu^{z+} ion mobility of many orders of magnitude higher compared to extrapolated high temperature values for bulk samples[22, 23].

Moreover, macroscopic room temperature insulators also used as high-k dielectrics e.g. Ta_2O_5, HfO_2, Al_2O_3 etc. demonstrate a sufficient ability to transport either oxygen ions or cations to ensure resistive switching (filament formation) within or below nanoseconds[11, 24-25]. These systems properties make possible using conventional electrochemical techniques e.g. cyclic voltammetry, pulsed techniques, chronoamperometry, chronopotentiometry, and steady state measurements[22, 26-28].

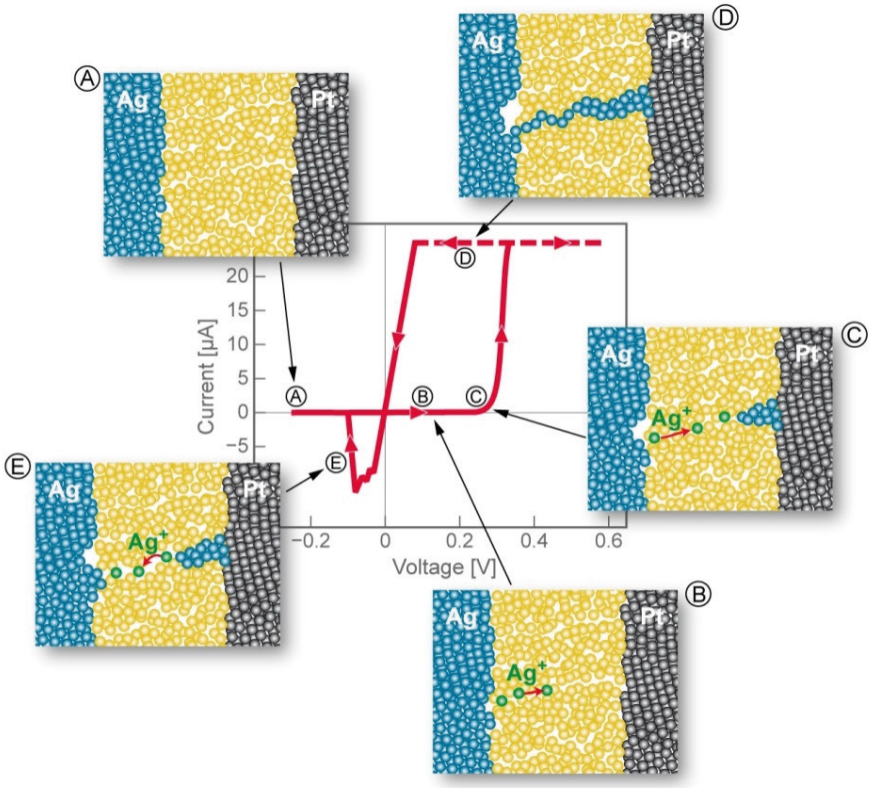


Fig. 2: *Schematic presentation of the processes during I-V sweep including formation and dissolution of a metallic filament in electrochemical metallization memories (ECM). The current saturation is due to the set current compliance (here 25 µA) used to prevent irreversible cell damages once the filament is formed. The value for the current compliance is chosen depending on the particular electrochemical system. It is also used to adjust the ON resistance (multi-level switching). The higher the current compliance is the lower is the ON resistance. Using different current compliances the ON resistance can be varied by orders of magnitude. The figure is adapted from reference [19]*

However, the systems' behavior also raises variety of theoretical and experimental difficulties: Simple calculation provides that to achieve so short switching times the diffusion coefficient must be in the range of 10^{-4} to 10^{-3} cm2 s^{-1}. The high diffusion constants cannot be explained in terms of chemical diffusion or high field drift but requires involving additional parameters like local Joule heating or moisture effects[27, 29-30]. In addition the charge screening (Debye) length for these classes materials is in the range of up to 100 nm i.e. much larger compared to the electrolyte thickness and thus, the condition for charge electroneutrality may not be fulfilled and space charge effects modulate the system properties [31-33].

The following contribution aims to give an overview on the current state-of-the-art understandings on the ECM cells and provide a discussion on the open questions and prospectives.

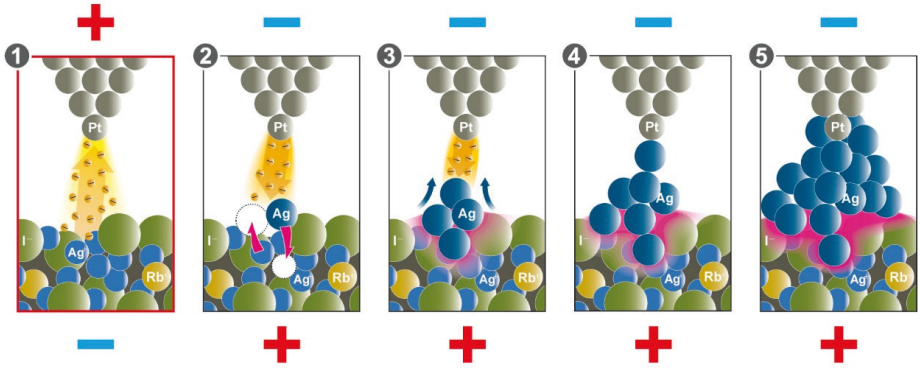


Fig. 3: *Schematic presentation of the functional principle of the gap type atomic switch. In a step (1) the surface is imaged at positive tip voltages of the STM. In (2) a negative voltage pulse is applied with intensity sufficiently high to overcome the reduction barrier for the Ag^+/Ag redox reaction. The filament starts to grow in (3) forming a single point contact with the STM tip (4). With time under the applied voltage the filament broadens reducing the cell resistance (5). The figure is adapted from reference [20]*

2 Materials

2.1 Solid Electrolyte Materials

Taking into account only the operation characteristics (switching time, retention, endurance, power consumption) of different types ReRAM cells, the particular materials used as solid electrolytes[33] appear at the first glance to be of a less importance. Looking to the vast number of publication in the literature the difference between using ion conductors (e.g $RbAg_4I_5$), mixed conductors (e.g. Ag_xS; Ag-doped GeS_x; Cu_xS, NiO) or insulators (e.g. SiO_2; Ta_2O_5, HfO_x) seems to be a question of device optimization. Due to the small diffusion lengths, the small amount of transferred mass, and high electric fields all these materials support to a sufficient extend ion movement to ensure the filament formation or dissolution within nanoseconds.

However, there are qualitative and quantitative differences using different classes of materials[33]. Some of the solid electrolytes are strongly stoichiometric e.g. $RbAg_4I_5$ and AgI. They can chemically dissolve neither the Ag electrode nor the formed tiny nano-filament. Moreover, the electrode reactions i.e. the dissolution and reduction of silver will not change their chemical composition throughout the film thickness, including layers adjacent to the electrode. In contrast, non-stoichiometric materials e.g. Ag_xS, Ag-doped GeS_x etc. exhibit compositional inhomogeneity and enrichment/depletion regions. Their transport properties (conductivity, ion and electron transference numbers) strongly depend on the chemical composition (or level of non-stoichiometry). Ag/GeS_x and Ag/$GeSe_x$ systems are examples for materials with time dependent composition and transport properties due to continuous dissolution of Ag until reaching the saturation limit[34-36]. For most of the cases the chemical activity of these electrolytes contributes to electrode and filament dissolution processes and thus, cell/device failures[28, 37].

Another materials class is represented by oxide insulators e.g SiO_2, Ta_2O_5 etc. being unable to chemically dissolve the active electrode. However, they do not initially contain mobile ions. In

this case in order to enable the electrode reaction of electrochemical oxidation (dissolution) of the active electrode one also necessarily needs a counter electrode reaction. It has been demonstrated that in many cases moisture (water reduction) is able to provide this counter electrode reaction thus, playing a crucial role for device operation and reliability[38-40]. Also protons and moisture have being considered as essential part of TiO_2 based VCM[41] and NiO based TCM devices[42].

Most important, these different types of materials can transit from one form to another.

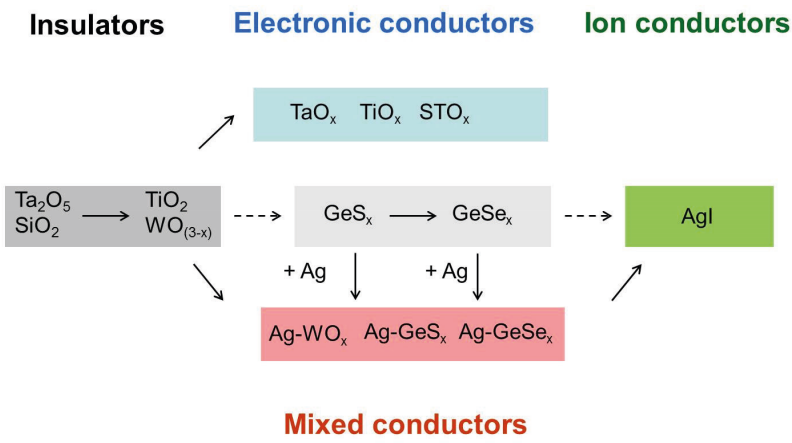


Fig. 4: *Transition of transport properties of ReRAM ion transporting solids due to chemical changes. The transition can occur due to FORM step or during SET and RESET operation or in some cases stimulated by UV, VIS radiation, moisture etc.*

As seen from the figure insulators can become electronic or mixed conductors or in particular cases e.g. strongly Ag-doped GeS_x and $GeSe_x$ to almost purely ion conductors[34, 43]. Also oxides such as WO_x, ZrO_2 and SiO_2 are reported to be pre-doped by thermal treatment to improve the switching characteristics[44-47]. Similar effect is also achieved by the FORM step in insulators where the materials are either reduced (VCM or TCM) or foreign ions are introduced (ECM). A sketch of the two possible situations is shown in Figure 4 on the example of ECM cells.

In both cases discussed in figure 4 the electrolyte changes its chemical, physical and transport properties showing compositional inhomogeneity and depletion/enrichment regions. In fact ReRAM solid electrolytes with nearly constant stoichiometry can be regarded only AgI, $RbAg_4I_5$ and Ag_2S.

Thus, the choice of solid electrolyte is determining the ReRAM cells stability and the reproducibility of the switching characteristics and the device performance in general.

It has to be mentioned that same materials can be used as electrolytes for all types ReRAM cells. The most typical example of this phenomenon is TiO_2 reported to be ECM material using e.g. Cu active electrode[48, 49], VCM material[50, 51], and TCM material[50, 52]. Many others e.g. Ta_2O_5, Al_2O_3, SiO_2, ZrO_2, HfO_2 etc. are reported to show dual or triple mode switching[13].

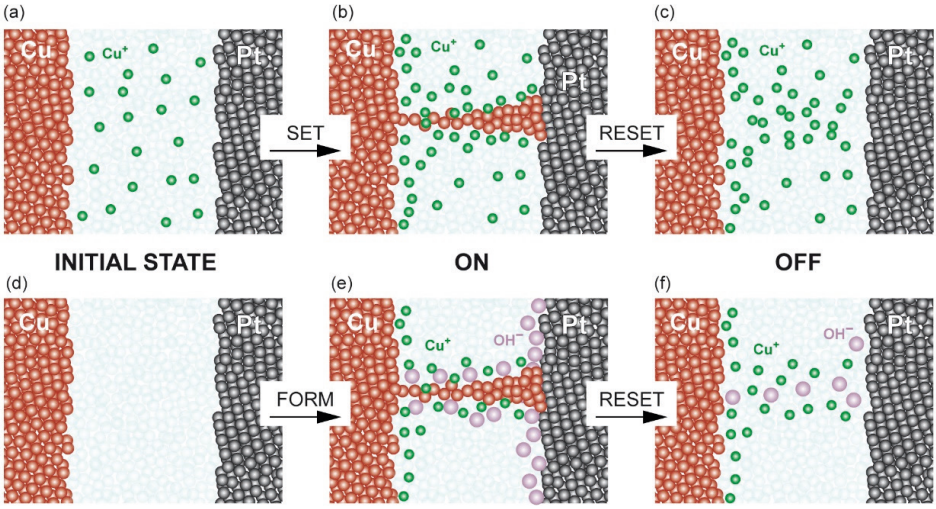


Fig. 5: *(a-c) Electrolytes initially containing mobile ionic species. (d-f) Insulators where the ionic species are incorporated during the FORM step. Due to the SET/RESET cycles of the cells, an inhomogeneous distribution of the mobile species is expected resulting in chemical potential gradients. The figure is reproduced from[33].*

2.2 Electrode Materials

The Active Electrode

As active electrode, typically Ag or Cu is used. Fig. 6 shows cyclic voltammetry (CV) experiments performed using Ta_2O_5 as solid electrolyte, in voltage range selected in a way to avoid resistive switching.

Several redox peaks were registered, each being an indication for electrochemical reaction. It can be seen that the shape of the CV and the number of peaks (i.e. redox reactions) depends on the active electrode material. Moreover, the reactions related to Ag proceed at lower voltage i.e. the following switching (not shown here) occurs at lower voltages. In the case of Cu electrode, one needs to apply roughly 2 V to switch the cell to LRS, whereas for Ag electrode this voltage is only 0.25 V.

The observed difference can be explained for general physicochemical considerations involving both thermodynamic and kinetic factors. Each CV shows the redox behaviour of different redox couple i.e. Ag^+/Ag and Cu^{x+}/Cu. The standard equilibrium potential for those couples as well as for many others can be found in the reference literature e.g. [54]. Taking three materials often used as active electrodes e.g. Ag, Cu and Ni the standard electrode potentials for the redox couples Ag^+/Ag, Cu^{2+}/Cu and Ni^{2+}/Ni (vs. standard hydrogen electrode) have the values 0.79 V; 0.34 V and -0.25 V, respectively. The physical meaning of these values is following: The lower the standard redox potential the higher the tendency to oxidation i.e. at positive applied voltages first will oxidize Ni, followed by Cu and Ag. For the reduction reaction, the situation is the opposite – the higher the redox potential the higher the tendency of reduction i.e. first will be reduced Ag^+, followed by Cu^{2+} and Ni^{2+}. One has to keep in mind that the standard electrode

potentials are thermodynamic quantity and often, kinetic restrictions (e.g. formation of passive films) may lead to deviation of expected properties. As an additional kinetic factor we also have to consider is the polarizability of the half-cell reactions i.e. the required overvoltage. Here Ag^+/Ag reaction has the lowest polarizability. All these factors i.e. equilibrium redox potentials, kinetic limitations and polarizability will play a role for the switching and must be considered. Most important new studies have shown that also metals such as Ta and Ti, typically used for the oxygen-based ReRAMs (VCM) can serve as active electrodes and contribute to the resistive switching by ECM mechanism[17, 18].

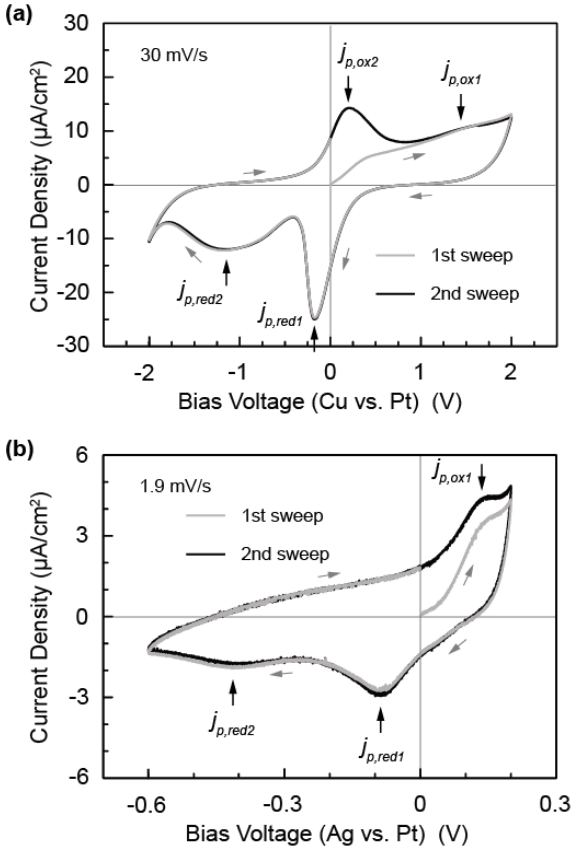


Fig. 6: *Cyclic voltammetry (or I-V sweeps) performed in the voltage range where no resistive switching is taking place in a) $Cu/Ta_2O_5/Pt$ cell and b) in $Ag/Ta_2O_5/Pt$ cell. The measurements were performed in a restricted voltage window in order to avoid the switching. The figure is reproduced from[53]*

In summary, the selection of active electrode material is an important prerequisite for proper operation of the cell and can significantly influence the cell kinetics.

The counter electrode and influence of moisture

The counter electrode material plays an important role in ReRAM cells. Its influence is twofold. First, it can catalyze or inhibit the reaction of the metal cation that forms the filament. Secondly, it can also catalyze or inhibit other parallel electrode reactions that can support or compete the main one e.g. redox reactions with moisture.

Influence of CE ion on the main cation redox reaction

To time, no particular studies have been reported on this issue. It is only worth mentioning that typically used counter electrodes in ECM/CBRAM cells are W, TiN, Pt, Ir, Ni, TiW, Au etc.[19, 33]. From more general point of view, it is expected that these different counter electrode materials will show different electro-catalytic activity towards the active metal redox reaction. To our best knowledge this issue has not been considered in the ReRAM literature and requires more attention.

Influence of CE on parallel redox reaction(s)

As discussed in section 2.1, despite apparent similarity between materials used as solid electrolytes to transport cations, one have to distinguish between materials that initially contain the mobile cations or such, that do not. Examples for the latter are SiO_2, Ta_2O_5, Al_2O_3 etc. To incorporate active metal ions within such materials a counter charge should be introduced to keep the electroneutrality[26]. For some type of solids this counter charge (or the related counter electrode reaction) can be electrons, i.e. reduction of the material, or can be absorbed moisture as in the case with Ta_2O_5[38] and SiO_2[26, 38]. In this situation in the entire ReRAM cell we have two electrode reactions – at the active metal electrode and at the counter electrode. The slower one will determine the overall reaction rate. In Fig. 7 are shown cyclic voltammetry experiments for ECM/CBRAM cells with different counter electrode materials, but using the same active electrode. It has to be noticed that without moisture the samples cannot be tested or formed at all, and show only irreversible hard breakdown.

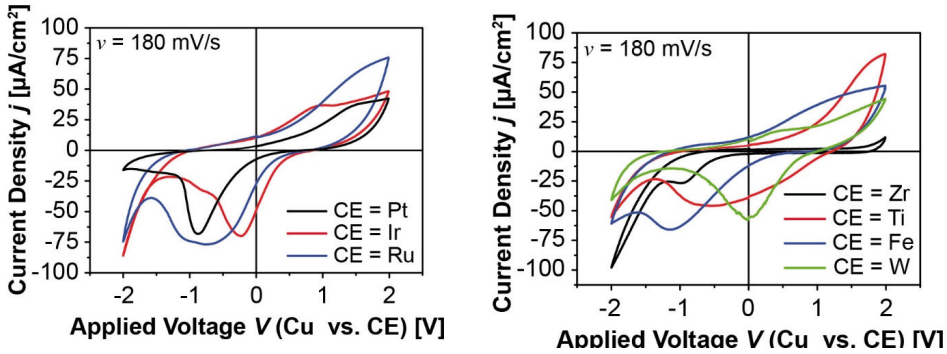


Fig. 7: *Cyclic voltammetry in SiO2-based cells using Cu active electrode and different counter electrode materials. The figure is reproduced from[55]*

As it can be seen, different counter electrode materials indeed show different electrocatalytic activity. Because the same active electrode (i.e. Cu) is used for the experiments, we concluded that slower reaction is the reaction at the counter electrode. Otherwise, if the active electrode reaction would be rate limiting, we would not have observed different behaviour for the different counter electrodes.

It has been verified by further experiments e.g. Infra-Red (IR) spectroscopy and electromotive force (emf) measurements, that exactly moisture in these materials is responsible to provide the required counter charge and counter electrode reaction.

Figure 8 shows the effect of the water partial pressure (pH_2O) on the properties of the cell.

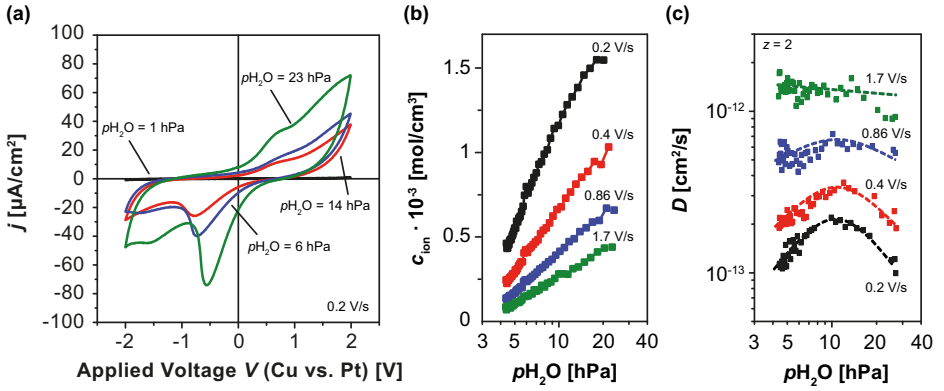


Fig. 8: *Influence of the pH_2O cell behavior of Cu/SiO_2/Pt cells. a) CV's at different pH_2O, b) Dependence of the incorporated ion s' concentration on the pH_2O and on the sweep rate. c) The influence of those both parameters on the estimated diffusion coefficient. The figure is modified from[26].*

The cyclic voltammograms showed increase in reaction rates (higher peak currents) for higher pH_2O. The calculated concentrations of the incorporated metal ions also is increased (Fig. 8b). However, the calculated diffusion coefficient shows a maximum, as it would be expected for high ion concentrations due to ion-ion interactions[26]. Effects of moisture on the resistive switching effects have been reported also for several other systems, for both ECM and VCM systems – Al/Al_2O_3[56], Pt/$SrTiO_{3-\delta}$[57], Pt/TiO_2[41], Pt/Ta_2O_5[18] and Ta/Ta_2O_5 systems[18]. Therefore, the influence of the counter electrode material is an essential parameter for determining the electrochemical reaction rate. However, the effects of moisture (both the positive and negative aspects) are not sufficiently well studied for most of the ReRAM systems and require additional attention.

2.3 The nanobattery effect

The nanobattery effect proves that ReRAMs behave as an active circuit element[58]. More important, it influences the device kinetics due to built-in voltage or alone by the chemical potential gradients i.e. concentration gradients within the cells. These gradients can appear in as deposited cells due inhomogeneous dissolution and retarded diffusion, or can be induced during SET/RESET operations due to unequal reaction rates of the reduction and oxidation processes of the same redox couple.

We discuss three factors which mainly contribute to the formation of the cell voltage V_{emf}, as shown in Fig. 9 for the example of a Ag/SiO_2/Pt cell – the diffusion potential V_d, the classical Nernst potential V_N and the potential due to the different surface free energies of macro- and nanoparticles V_{GT} for the case that a metallic nanofilament forms or is incompletely formed.

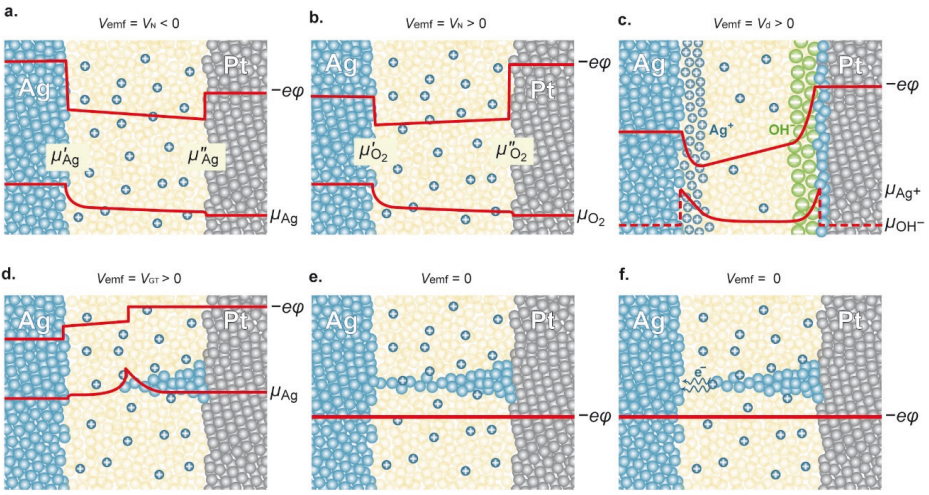


Fig. 9: *Origins of emf in nanoscale ECM cells. (a) Gradient of the chemical potential of Ag metal at the interfaces Ag/electrolyte and Pt/electrolyte* $\Delta\mu_{Ag} = \mu'_{Ag} - \mu''_{Ag}$. *The* V_{emf} *as given by equation (9) has a negative value and is then given by the difference of the electrical potentials at both electrodes* $\Delta\varphi = V_{emf} = -\frac{\Delta\mu_{Ag}}{ze}$ *generated to keep the condition* $\sum \tilde{\mu}_i = 0$ *(*$\tilde{\mu}_i$ *is the electrochemical potential given by* $\tilde{\mu}_i = \mu_i + ze\varphi$*). (b) Gradient of the chemical potential of molecular oxygen at the interfaces Ag/electrolyte and Pt/electrolyte* $\Delta\mu_{O_2} = \mu'_{O_2} - \mu''_{O_2}$ *(pO2 is defined by the ambient atmosphere). The Vemf is given by equation (9) and has in this case a positive value. (c) The emf is generated by gradients of the chemical potentials of the* Ag^+ *and* OH^- *ions i.e.,* $\Delta\mu_{Ag^+} = \mu'_{Ag^+} - \mu''_{Ag^+}$ *and* $\Delta\mu_{OH^-} = \mu'_{OH^-} - \mu''_{OH^-}$ *inhomogeneously distributed in the thin film and an additional term due to the different half-cell reactions as given by equations (7) and (8). (d) In the case of a nanosize filament, the chemical potential of Ag contains an additional surface energy term generating a chemical potential gradient* $\Delta\mu_{Ag} = \mu_{Ag\text{-micro}} - \mu_{Ag\text{-nano}}$ *in accordance with equation (10). In the case of fully metallic contact or tunnel junction, the emf is* $V_{emf} = 0$ *(e, f). The potential of the Pt electrode is used as a reference. The figure is adapted from[58].*

The nonequilibrium diffusion potential V_d in ECM cells (not observable for e.g. AgI-based systems) compensates the chemical potential gradient arising due to the inhomogeneous distribution of charged species, i.e., Ag^+ ions, electrons and/or OH^- ions within the electrolyte film. These ions are introduced into the solid films either electrochemically during the forming of SET/RESET cycles or chemically due to a chemical dissolution of Ag. In both cases, the metal/nonmetal ratio changes across the electrolyte layer, with particularly pronounced changes in the vicinity of the electrodes. The electromotive force generated by this inhomogeneous charge distribution and mobilities is given by[59]:

$$V_d = -\frac{kT}{e}\sum_i \int_{s''}^{s'} \frac{t_i}{z_i} d\ln a_i = -\frac{kT}{e}\left(\bar{t}_{Me^+}\frac{(a_{Me^+})_{s'}}{(a_{Me^+})_{s''}} - \bar{t}_-\frac{(a_-)_{s'}}{(a_-)_{s''}}\right) \quad (1)$$

where k, T and e are the Boltzmann constant, the temperature, and the elementary charge, respectively; t_i denotes the transference number of the species i, z_i the charge of this species and

a_i their activities at the interfaces s' and s" corresponding to active electrode/electrolyte and inert electrode/electrolyte interfaces, respectively.

The potential difference generated in the neuron cells to transport electric signals has the same nature and originates in the diffusion (Donnan) potential[59].

The Nernst voltage is given by the difference between the potential-determining half-cell reactions at each electrode/electrolyte interface:

$$V_N = V_{s'} - V_{s''} = V^0 + \frac{kT}{e}\ln\frac{(a_{Me^{z+}})_{s'}\cdot(a_{Red})_{s''}}{(a_{Me})_{s'}\cdot(a_{Ox})_{s''}} \tag{2}$$

with $V_{s'}$ and $V_{s''}$ being the half-cell potentials at the active electrode/electrolyte (s') and inert electrode/electrolyte (s") interfaces, and V^0 is the difference in the standard potentials of these reactions.

At the s' interface, the potential-determining reaction is the same for all ECM cells:

$$Me^{z+} + ze^- \rightleftharpoons Me \tag{3}$$

At the s" interface, the potential-determining reaction depends on the material properties of the solid electrolyte and we distinguish two boundary conditions. For the first boundary condition the solid film contains no Me^{z+} and reaction (3) cannot be potential-determining. Instead, moisture is the crucial factor in providing the required counter reaction at the inert electrode[38] e.g.,

$$\tfrac{1}{2}O_2 + H_2O + 2e^- \rightleftharpoons 2OH^- \quad \left(\text{or alternatively } 2H_2O + 2e^- \rightleftharpoons 2OH^- + H_2\right) \tag{4}$$

and the Nernst voltage takes the form:

$$V_N = V^0 + \frac{kT}{2e}\ln\frac{(a^2_{Me^{z+}})_{s'}\cdot(a^2_{OH^-})_{s''}}{(a^2_{Me})_{s'}\cdot(a^{1/2}_{O_2})_{s''}\cdot(a_{H_2O})_{s''}} \tag{5}$$

The total emf of the ECM cell is a combination of equations (1) and (5) and is expressed by:

$$V_{emf} = V_N + V_d = V_0 + \bar{t}_{OH^-}\frac{kT}{e}\ln(a_{Me^+})_{s'} + \bar{t}_{Me^+}\frac{kT}{e}\ln(a_{OH^-})_{s''} \tag{6}$$

Here $V_0 = V^0$ + const. The emf for Ag/SiO_2/Pt cell is shown in Fig. 2a,b where this situation is most clearly observed. The amount of ions generated at the s' and s" interfaces during cell operation is adjusted by the pulse length and height or the sweep rate[27].

In the second boundary condition, the electrolyte contains mobile Me^{z+} ions dissolved by electrochemical and/or chemical processes with almost homogeneous distribution i.e., $(a_{Me^{z+}})_{s'} \sim (a_{Me^{z+}})_{s''}$. Then, at the s" interface, the potential-determining half-cell reaction will be the same as at s' i.e., reaction (3), and, because no chemical potential gradient of the charged species is assumed, equation (2) can be simplified to:

$$V_{emf} = \bar{t}_{ion}\frac{kT}{e}\ln\frac{(a_{Me})_{s''}}{(a_{Me})_{s'}} \tag{7}$$

where aMe is the activity of the active metal at both interfaces and $\bar{t}_{ion}$ is the ion transference number averaged throughout the electrolyte thickness. Because $(a_{Me})_{s''}$ is fixed, the Nernst voltage is a function of $(a_{Me})_{s''}$ alone. We succeeded in clearly demonstrating a Nernst emf in accordance with equation (9) (Fig. 1a) for the systems Ag/Ag-$GeS_{2.2}$/Pt and Ag/Ag-$GeSe_{2.3}$/Pt. Initially, we applied a positive external voltage to the inert electrode, thus removing the residual

Ag atoms from the s" interface. In the related cathodic reaction, Ag whiskers are deposited at the Ag electrode as shown in Fig. 2c. The emf of the system was then monitored over time and correlated to the evolution of the whisker morphology recorded by optical images. In the presence of the Ag whiskers formed, the emf value decreases to –450 mV in accordance with equation (7) corresponding to a difference $(a_{Ag})_{s'} = 4 \cdot 10^7 (a_{Ag})_{s''}$. As the whiskers begin to dissolve the emf increases simultaneously again to a positive relaxation voltage by a change in the control of V_{emf} from a situation determined by equation (7) to a situation determined by equation (1).

The contribution of the third component V_{GT} to the total cell voltage V_{emf} is expected in the case of a noncontacting filament due to the different surface free energies of the macrocrystalline active electrode and the nanosize filament in accordance with the Gibbs-Thomson equation:

$$V_{GT} = -\frac{\mu_{Ag}^{macro} - \mu_{Ag}^{nano}}{ze} = -\frac{2\gamma}{zer} V_m \tag{8}$$

here γ is the surface free energy, r is the radius of the particle, V_m is the molar volume and μ_{Ag} is the chemical potential of Ag. V_{GT} can be observed either for highly ohmic ($R > 12.9$ kΩ) ON states (no metallic short circuit) or for thin filaments where the loss of a few atoms breaks the metallic contact. Thus, V_{GT} leads to a complete loss of the ON state. The discussed chemical potential gradient(s) are also present and contribute to the spontaneous dissolution of the nanofilament. However, the contribution of V_{GT} could not be clearly distinguished from the contribution of V_d.

Thus, we were able to prove that in all types of ReRAM cells tested, significant chemical potential gradients are generated in both a chemical and electrochemical manner by the operation of the cells. Inevitably, these gradients give rise to emf, and, hence, the cells show the characteristics of nanosize batteries.

The non-equilibrium states affect both the retention and the device operation. It influences the device kinetics due to built-in voltage or alone by the chemical potential gradients i.e. concentration gradients within the cells. These gradients can appear in as deposited cells due inhomogeneous dissolution and retarded diffusion, or can be induced during SET/RESET operations due to unequal reaction rates of the reduction and oxidation processes of the same redox couple. In fact, the voltage measurement is only a tool to detect the gradients. It is important to note, that often due to higher electronic partial conductivity of the solid electrolytes the measured voltage (due to the nanobattery effect) is much smaller compared to the expected theoretical value that corresponds to the concentration gradients. Eq. 9 shows this dependence.

$$V_{cell} = \overline{t}_{ion} V_{emf} \tag{9}$$

V_{cell} is the measured voltage on the device, t_{ion} is the average transference number of the ions (i.e. the part of the current transported by ions) and V_{emf} is the theoretical value expected for the nanobattery effect. In many electronic oxides used for resistive switching the transference number of the ions is as low as 10^{-3} or even less[58].

For example, ion ($z = 1$) concentration gradient of 10^{17} will cause $V_{emf} \sim 1$ V. Using solid electrolyte with a t_{ion} of 10^{-3} we will measure V_{cell} of only 1 mV. However, the concertation difference of 10^{17} is persistent and will impact the cell behavior.

Figure 6 shows simulations on the influence of the nanobattery effect on SET kinetics (voltage).

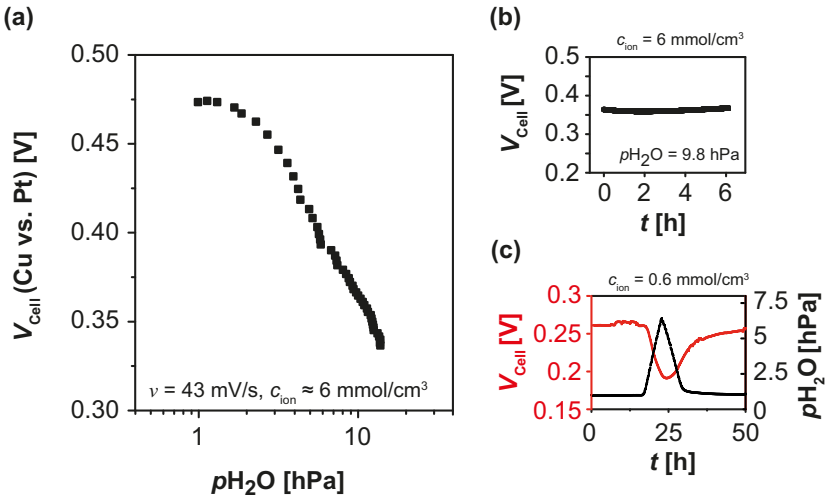


Fig. 10: *Electromotive force measurement at a constant ion concentration. (a) V_{Cell} dependence on the water partial pressure in nitrogen atmosphere. Each point corresponds to a different pH_2O value. The initial ion concentration (indicated in the plot) was set prior to the measurements. (b) Transient measurement of V_{Cell} at constant c_{ion} and pH_2O. (c) V_{Cell} (red) response to the variation of the pH_2O (black). The V_{Cell} response on pH_2O change is reversible. This indicates that the concentration gradient of Cu remains constant with time and pH_2O does not meaningfully increase the mobility of Cu ions in the SiO_2 . The figure is reproduced from[26].*

The nanobattery effect was also found to be strongly influenced by the local environment. As an example, the cell voltage of the cell Cu/SiO_2/Pt has been measured at different water partial pressures pH_2O in nitrogen atmosphere as depicted in Figure 10a. cion was adjusted prior to the experiment by a single linear anodic oxidation sweep. The ion concentration of both cations and OH^- is believed to be constant during the experiment. From eq. (5) it is evident that the emf is influenced by the c (H_2O) (and thus, pH_2O) also when c (Cu^{x+}) and c (OH^-) are constant. To ensure reproducible experimental conditions, the oxygen partial pressure (pO_2) was monitored by an oxygen sensor simultaneously and was found to be constant during the measurements. The highest value for V_{cell} is measured in anhydrous nitrogen atmosphere. When pH_2O is increased a decrease of V_{cell} is observed in accordance to equation (5). The system requires at least 60 min relaxation as soon as quasi-equilibrium for each partial pressure of water is reached. Figure 10b shows that V_{cell} remains constant over hours. V_{cell} can be reversibly tuned (increased or decreased) depending on the particular water partial pressure as depicted in Figure 10c. Hence, pH_2O and therefore, the water molecules incorporated into SiO_2 do not significantly increase the ion mobility (due to solvent effects) and the Cu^{z+} and OH^- ion concentration gradients and respectively, the driving force for V_{cell} are not changed.

We assume moisture is likely penetrating from lateral sides. However, resistive switching experiments[38, 60] indicate a pH_2O equilibration in the complete SiO_2 thin film thus, even underneath the top electrodes. This complies with our observation that the V_{Cell} equilibration can take up to several hour depending on the change of ambient water partial pressure.

3 Processes

3.1 Switching Kinetics

Filament formation

The kinetics of the filament formation has been intensively studied in the recent years[14-15, 29-30, 33, 61-64] and several factors e.g. Joule heating, temperature gradients, chemical potential gradients, ambient moisture etc. were found to play an important role[26, 28, 30, 33, 38, 65]. Depending on the different types of ReRAM cells nucleation[20, 24, 66-69], charge transfer[70, 71] or ion drift[72] were reported to be rate limiting. The origin of the rate limiting step however, cannot be unified because of its dependence on the nature of the solid electrolyte, on the electrode material(s) and on the thermodynamic factors. There are only two generalized kinetic models offered in the literature combining the essential system-relevant factors and suggesting the conditions where one or other process will become rate limiting[29, 30]. As shown in Fig. 11 different regimes' representation depends on the applied voltage and temperature.

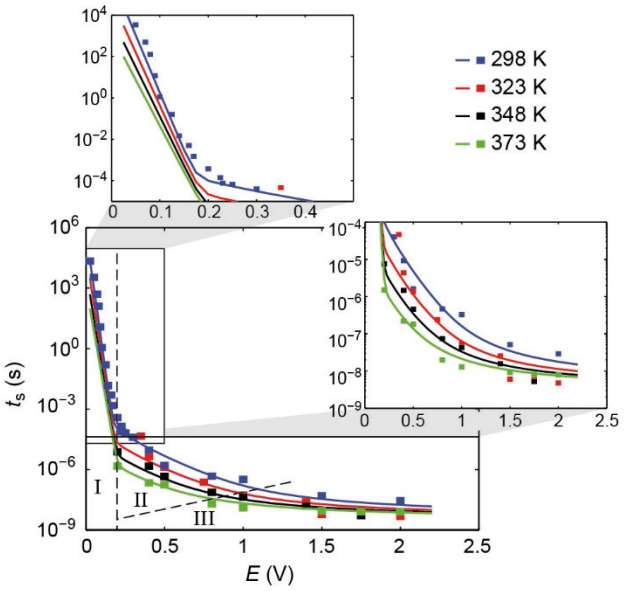


Fig. 11: *Pulsed SET switching kinetics of the AgI-based ECM cell for different ambient temperatures* T = *298 K (blue), 323 K (red), 348 K (black) and 373 K (green). The simulated data is displayed using solid lines and the experimental data using squares. I, II, III mark the nucleation limited, the electron transfer limited and the mixed control regime, respectively. The figure is reproduced from reference [29].*

The switching time t_s is then a sum of the times required for individual processes. In regime I at lower voltages the nucleation is rate-limiting i.e. the time t_n to form critical nucleus.

$$t_n = t_0 \exp\left(\frac{\Delta G_n^{\neq}}{kT}\right)\exp\left(-\frac{(N_c+\alpha)ze}{kT}\Delta\varphi\right) \tag{10}$$

with t_0 being a pre-exponential factor[20]; $\Delta G_n^{\neq}$ is the activation energy for the process (including the excess surface energy term); N_c is the number of atoms constituting the critical nucleus; α is the transfer coefficient and $\Delta\varphi$ is the applied voltage. For this conditions $t_s = t_n$.

Increasing the applied voltage lowers the effective nucleation free energy and the charge transfer process given by the Butler-Volmer equation becomes rate limiting (region II):

$$j = j_0 \left[\exp\left(\frac{(1-\alpha) ze}{kT} \Delta\varphi \right) - \exp\left(-\frac{\alpha ze}{kT} \Delta\varphi \right) \right] \tag{11}$$

where j and j_0 are the current density and the exchange current density, respectively. Thus, the time t_e for the charge transfer is determining the switching time i.e. $t_s = t_e$. Finally at much higher applied potentials (as in region III) the process of diffusion of ions within the solid film given by the diffusion current will limit the switching time ($t_s = t_d$):

$$j_d = 2zecaf \exp\left(-\frac{\Delta G_d^{\neq}}{kT} \right) \sinh\left(\frac{aze}{2kT} E \right) \tag{12}$$

with j_d denoting the diffusion determined current density, a is the mean ion jump distance, f the attempt frequency, $\Delta G^{\neq}_d$ the jump activation barrier height, c the ion concentration, and E the applied electric field.

The suggested model has been verified within 14 orders of magnitude variation of the switching time on the system Ag/AgI/Pt but can be applied to any material system. However, the particular distribution of regimes I, II and III will individually vary.

Filament dissolution

Not much is known on the kinetics limitations and the mechanism of the RESET process and only few papers report on it[73, 74].

Two different regimes have been identified: a low voltage regime up to roughly −200 mV and a high voltage regime from −1 V and higher. For the low voltage regime, external effects may influence the cell behavior, such as thermal effects or the nanobattery effect. Both can RESET the memory cell spontaneously[75]. Thus, the low voltage RESET times (on the example with Ag-GeS system[74]) itself are most probably slightly higher compared to the measurements.

The experimentally observed RESET behaviour can also be explained within the frame of the same simulation model used for the SET kinetics. In the low voltage regime described above, the RESET process is limited mainly by the charge (electron) transfer reactions. With increasing the applied voltage, ion migration becomes substantial for the RESET kinetics and finally limits the RESET process. The proposed model is based on the assumption that the filament is completely dissolved during the RESET process. The latter has been also experimentally verified[74].

3.2 Filament dynamics and quantum point contacts

Filament shape and growth direction

First reports on direct observation and visualization of a filament were on the atomic switch[12]. Later on, reports on different systems appeared showing also the filament dynamics i.e. its possible stabilization, growth or dissolution in atomic switch configuration[8, 20].

In ECM/CBRAM devices, the conditions for tracking the filament are unfavorable and several difficulties have to be overcome. One of the main problems is succeeding to avoid planar devices (which are more easy to prepare and investigate), because the conditions for ion and atom diffusion, as well as the defect states at the surface differ significantly from those in the volume. In addition, in the volume of the film, there are space restrictions and the formation of a new

phase (with different lattice parameters) is related with some mechanical stress (mainly compressive), whereas this problem is not relevant for surfaces. The first report on *in situ* TEM study on ECM/CBRAM cells of the type Ag/Ag_2S/W was by Xu et al.[76]. Further studies followed, going into details in the switching mechanism and the filament dynamics[77-79]. A very interesting approach using 3D tomography has been recently used by Celano et al.[80, 81] demonstrating the applicability of this method to resistive switching memories.

According to the initial concept, the filament forms at the counter electrode (negatively biased) during the SET process. The shape was approximated to either cylindrical or conical one with a tip pointing to the active electrode[19, 82]. An additional inversed growth mode has been reported, showing that the filament can also grow from the anode towards the cathode[77-79, 83]. Different explanations have been offered and controversially discussed to explain the inversed filament growth mode [77, 83-88]. A recent study by Yang et al. have succeeded not only to explain microscopically the inversed mode, but also to classify the different switching mechanisms and formulate an unified framework[79]. In specially designed sample holder, the formation and dynamics of the filaments could be observed by *in situ* TEM within the volume of the solid film.

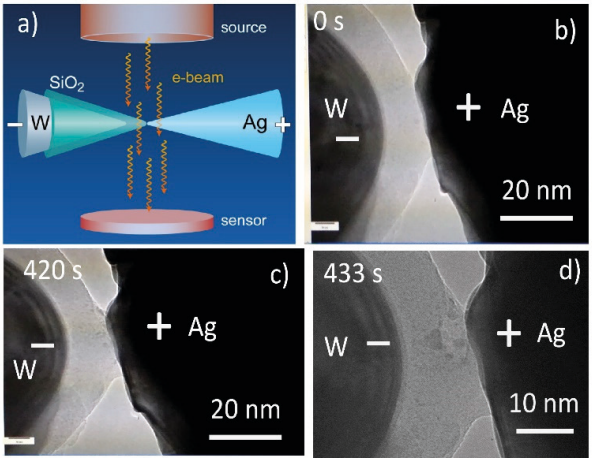


Fig. 12: *. In situ TEM experiments. a) Schematic presentation of the experimental SETup; b) image of W/SiO_2/Ag cell before voltage application and c-d) images at applied voltage of 8 V after 420s and 433 s, respectively. The figure is adapted from[79]*

It can be seen from Figure 12 that the filament expands from the anode towards the cathode. Detailed diffraction has shown that the formed particles are entirely metallic i.e. we cannot presume that Ag^+-ions have been incorporated into SiO_2, forming a ternary compound with increased electronic conductivity, but only Ag metal nanoparticles were detected. It has also been found that Ag from the electrode is consumed for the process and in a certain time the Ag-SiO_2 contact was lost. However, applying higher voltage allows to reveal it.

Further, it has been observed that voids remain at the position where Ag-nanoparticles were situated. In the same way as it grows the filament can be retracted back by applying the opposite (negative) voltage to the Ag electrode. The inversed growth mode was explained with the effect of the bipolar electrode[79, 87].

The different growth modes observed by several research groups can be all explained taking into account some kinetic factors, that is, electrode reaction rates, ion concentration within the

solid electrolyte and their distribution and the related mobility/conductivity. Four situations can be distinguished:

- The ion mobility and the redox rates are homogeneous and high. For this situation the dissolved ions can reach the inert electrode without being reduced and/or agglomerating, thus avoiding nucleation within the insulating film. The filament starts growing from the inert electrode, and the sufficient ion supply leads to conically shaped filament with a tip pointing to the active electrode. This situation describes the filament growth in conventional ECM cells, and corresponds to material systems with high ionic conductivity.
- The ion mobility is low and the electrode reaction rates are low. For these conditions the ions can pile within the solid electrolyte and reach the critical nucleation conditions and further filament can proceed by cluster displacement due to the bipolar electrode effect. An experimental example is the filament growth in amorphous Si, where the filament is initiated from the active electrode and grows towards the inert electrode as discrete nanoclusters[78]. This situation corresponds to material systems with very low ionic conductivity such as SiO_2 or Si.
- The ion mobility is low but the electrode reaction rates are high. Nucleation can now occur inside the dielectric while large amounts of atoms can be deposited onto the cathode sides of the nuclei, leading to gap filling. After a connection between the nuclei and the active electrode is achieved, the process is repeated leading to an effective forward growth towards the inert electrode.
- The ion mobility is high, while the redox reaction rates are low. For this situation nucleation only occurs at the counter electrode and the reduction predominately occurs at the edges (high field strengths), thus leading to branched filament growth towards the active electrode.

Based on this framework the different experimental observations can be fully accounted. In general, the ion mobility in a given dielectric determines the nucleation sites and the direction of the filament growth, while the redox rates determine the ion supply and the geometry of the filament.

Formation of quantum point contacts

The small size of the filament in the range of few nanometers in a diameter indicates that quantum effects in the conductivity have to be expected. First quantized conduction has been reported for gap-type atomic switches[12]. The single atomic point contact (Landauer) conductivity is given by $G_0 = 2e^2/h$ or the corresponding resistance of 12.9 kΩ.

In the resent years quantized conductivity has been also reported also for gapless-type switches[89-93]. The difficulty for detecting the quantum steps in gapless type cells is that the experimental conditions are restricted. If the applied voltage is too low no switching occurs. If it is too high then the switching is much faster than the time resolution of the equipment. Only in a small window of applied voltages (typically 20 mV to 50 mV) the filament growth and the formation of atomic point contacts (as shown in Figs. 2 or 3) can be detected. As alternative technique slow current sweeps can be used as shown in Fig. 14 to observe quantum steps.

Here, multiple integers of G_0 were observed during the current sweep dispersed with a certain statistical probability. The multiple quantum steps are discussed to serve for multibit memory storage but despite the attractive advantages of this opportunity to stabilize the quantum steps at room temperature is not an easy task, because the tiny nanofilaments are subject to diferent

chamical and electrochemical interactions (e.g. see Eq. 8) and easily dissolved. Thus, resistances in the range of some kOhms (corresponding to up to 10 G_0) were found unstable and increase with time. Higher number of multiple integers of G_0 is practically not convenient because the difference between the resistance levels is very small and difficult to distinguish.

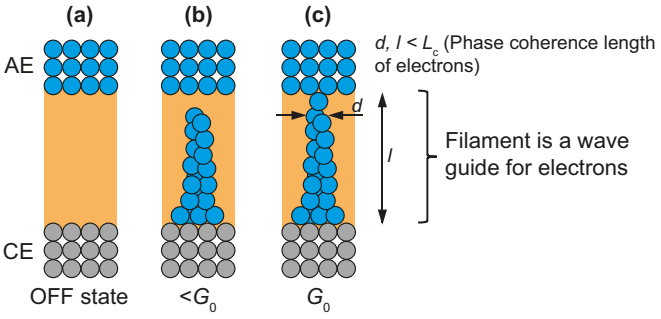


Fig. 13: *(a)-(c) Schematic presentation of an ECM ReRAM cell. In (a) the cell is in the high resistive OFF state. A filament is forming and tunneling dominates the conductance in (b) with $G < G_0$. In (c) one atom forms a metallic point contact between the AE and CE.*

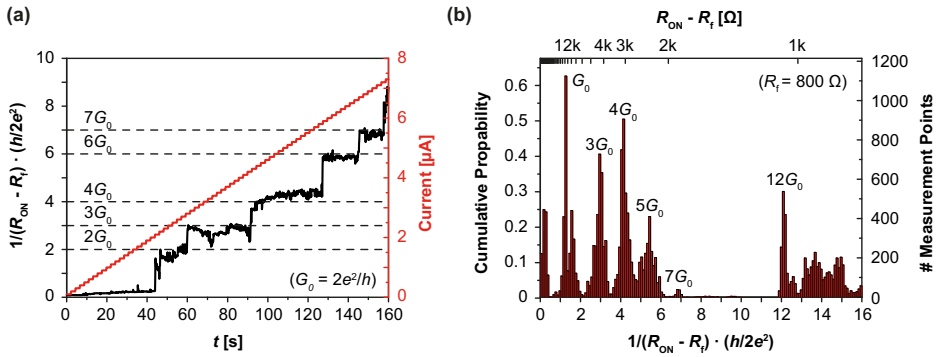


Fig. 14: *Analysis of quantized cell conductance. (a) At least five quantized resistances have been observed in this example by current sweeping. (b) Cumulative statistics of measured cell conductivity. The figure is reproduced from reference [89]*

Apart from its practical use the value of G_0 (or the resistance of 12.9 kΩ) serves as an important criterion whether a metallic contact has been established or not. ON resistances of ReRAM cells much higher than 12.9 kΩ is unlikely to be determined by a complete metallic filament. A calculation of the filament diameter/resistance dependence is shown in Fig. 15 for several electrolyte thicknesses.

In order to have a ON resistance higher than ~ 15 kΩ the filament should have a diameter smaller than an atomic diameter, which is physically impossible. Therefore higher ON resistances are usually interpreted in terms of a tunnel resistance.

In considering an appropriate ON-state resistance model, there is one important issue which is often overlooked or at least rarely discussed – the particular experimental procedure for determining the resistance. Often resistance is extracted from the linear slope of *I-V* sweeps immediately after the program or set process for different test conditions and parameters. In other cases R_{ON} is determined using peak current flow as part of endurance tests which use voltage

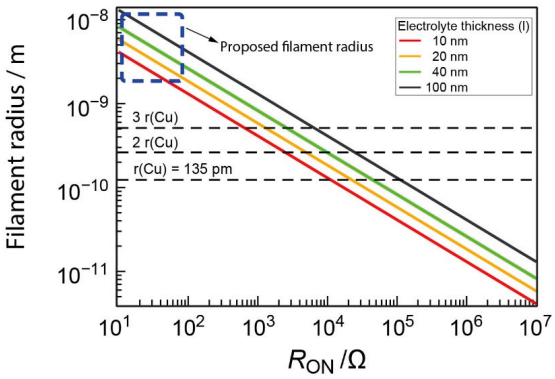


Fig. 15: . *Calculated ON state resistance of ECM type of cell as a function of the filament radius for solid electrolytes film thicknesses between 10 nm and 100 nm. The figure is adapted from[33].*

pulse or sweep cycling. The problem is that these methods are dynamical and may not give the cell time to reach a final stable (relaxed) state. For example, even though the initial switching can be extremely fast, in the order of ns, a further drop in resistance is often evident if the programming voltage is maintained across the cell [94], an effect that is thought to be due to a thickening of the initial bridging connection. Worse still, at low current compliances it has been shown that ECM cells have high ON-state resistance levels (in the high kΩ or MΩ range) extracted from the *I-V* sweeps and this resistance can be reproduced for many repeated cycles. However, if one stops the cycling of the cell in the ON-state and measures the resistance after some time one finds that the cell is already in the OFF-state, i.e., the high resistance ON-states exist only when the voltage (current) is applied and disappear after the system is left to equilibrate. These effects can lead to erroneous ON-state resistance measurement, particularly for resistances much greater than 13 kΩ.

As an additional proof on the nature of the ON state can be used the emf value which for metallic or tunnel contacts must be $E = 0$ V (see Fig. 9e,f). The quantum contacts can be broken due to influence of the environment e.g rest oxygen, moisture, temperature but also due to chemical dissolution within the electrolyte matrix.

4 Conclusions

Electrochemical metallization resistive switching memory cells are shown to be a representative example for solid state electrochemical systems with nano or atomic dimensions. The small size of the cells leads to unconventional effects and conditions of operation e.g. sufficient conductivity to carry out electrochemical measurements using thin films of bulk insulators, quantum size effects, etc. The issues in the materials selection for solid electrolytes, electrodes and the nanobattery effect are discussed. The influence of the local environment and in that sense of the moister has been pointed out. The electrochemical kinetics of filament formation and dissolution have been highlighted as well as the microscopic studies on the filament dynamics and formation of quantum point contacts.

References

[1] R. Waser and M. Aono, Nat. Mater. **6**, 833 (2007).

[2] I. G. Baek et al., IEDM, 587 (2004).

[3] Y. Yang, S. Choi, and W. Lu, Nano Letters **13**, 2908 (2013).

[4] J. Borghetti et al., Nature **464**, 873 (2010).

[5] S. Kaeriyama et al., IEEE Journal of Solid-State Circuits, USA **40**, 168 (2005).

[6] D. B. Strukov and K. K. Likharev, Nanotechnology **16**, 888 (2005).

[7] Y. V. Pershin and M. Di Ventra, IEEE Transactions on Circuits and Systems **57**, 1857 (2010).

[8] T. Ohno et al., Nat. Mater. **10**, 591 (2011).

[9] S. H. Jo et al., Nano Lett. **10**, 1297 (2010).

[10] M. D. Pickett, G. Medeiros-Ribeiro, and R. S. Williams, Nat. Mater. **12**, 114 (2013).

[11] A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, Nanotechnology **22**, 485203 (2011).

[12] K. Terabe, T. Hasegawa, T. Nakayama, and M. Aono, Nature **433**, 47 (2005).

[13] I. Valov, ChemElectroChem **1**, 26 (2014).

[14] T. Hasegawa, K. Terabe, T. Tsuruoka, and M. Aono, Adv. Mater. **24**, 252 (2012).

[15] J. J. Yang, D. B. Strukov, and D. R. Stewart, Nat. Nanotechnol. **8**, 13 (2013).

[16] R. Waser (Ed.), *Nanoelectronics and Information Technology*, Wiley-VCH, 2012.

[17] A. Wedig et al., Nat. Nanotechnol. (2015).

[18] M. Lübben et al., Adv. Mater. **27**, 6202 (2015).

[19] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, Nanotechnology **22**, 254003/1 (2011).

[20] I. Valov et al., Nat. Mater. **11**, 530 (2012).

[21] B. Govoreanu et al., IEDM Tech. Dig., 31.6.1 (2011).

[22] S. Tappertzhofen, S. Menzel, I. Valov, and R. Waser, Appl. Phys. Lett. **99**, 203103/1 (2011).

[23] D. Y. Cho, S. Tappertzhofen, R. Waser, and I. Valov, Nanoscale **5**, 1781 (2013).

[24] T. Tsuruoka et al., AIP Adv. **3**, 32114/1 (2013).

[25] H.- S. Lee et al., Phys. Rev. B: Condens. Matter **83**, 104110/1 (2011).

[26] S. Tappertzhofen et al., ACS Nano **7**, 6396 (2013).

[27] S. Tappertzhofen, H. Mündelein, I. Valov, and R. Waser, Nanoscale **4**, 3040 (2012).

[28] I. Valov et al., Nature Communications **4**, 1771 (2013).

[29] S. Menzel, S. Tappertzhofen, R. Waser, and I. Valov, PCCP **15**, 6945 (2013).

[30] S. Menzel et al., Adv. Funct. Mater. **21**, 4487 (2011).

[31] J. Maier, Adv. Mater. **21**, 2571 (2009).

[32] J. Maier, Nat. Mater. **4**, 805 (2005).

[33] I. Valov and M. N. Kozicki, J. Phys. D Appl. Phys. **46**, 074005 (2013).

[34] M. A. Urena, A. A. Piarristeguy, M. Fontana, and B. Arcondo, Solid State Ionics, Diffusion & Reactions, Netherlands **176**, 505 (2005).

[35] Y. Kawamoto, N. Nagura, and S. Tsuchihashi, J. Am. Ceram. Soc. **57**, 489 (1974).

[36] C. McHardy, A. Fitzgerald, P. Moir, and M. Flynn, J. Phys. C Solid State Phys. **20**, 4055 (1987).

[37] D.-Y. Cho et al., Advanced Materials **24**, 4552 (2012).

[38] T. Tsuruoka et al., Advanced Functional Materials **22**, 70 (2012).

[39] X. Yang, B. J. Choi, A. B. K. Chen, and I. W. Chen, ACS Nano **7**, 2302 (2013).

[40] B. J. Choi and I-W. Chen, Appl. Phys. A Mater. Sci. Process. **112**, 235 (2013).

[41] J. R. Jameson and Y. Nishi, Integrated Ferroelectrics **124**, 112 (2011).

[42] W.-L. Jang, Y.-M. Lu, and W.-S. Hwang, Vacuum, 1 (2008).

[43] M. Kawasaki, J. Kawamura, Y. Nakamura, and M. Aniya, Solid State Ionics **123**, 259 (1999).

[44] C. Gopalan et al., J. Non-Cryst. Solids **353**, 1844 (2007).

[45] S. P. Thermadam et al., Thin Solid Films **518**, 3293 (2010).

[46] C. Schindler, S. C. P. Thermadam, R. Waser, and M. N. Kozicki, IEEE Trans. Electron Devices **54**, 2762 (2007).

[47] Q. Liu et al., Appl. Phys. Lett. **95**, 23501/1 (2009).

[48] L. Yang et al., Appl. Phys. Lett. **95**, 013109 (2009).

[49] Y. Huang, H. Lin, and H. Cheng, IEEE Electron Device Lett., 236 (2013).

[50] S. Kim et al., J. Mater. Res. **28**, 313 (2013).

[51] K. Szot et al., Nanotechnology **22**, 254001/1 (2011).

[52] D. Ielmini, R. Bruchhaus, and R. Waser, Phase Transit. **84**, 570 (2011).

[53] T. Tsuruoka et al., Adv. Func. Mat. **25**, 6374 (2015).

[54] W. M. Haynes, *CRC Handbook of Chemistry and Physics*, CRC press, 2013.

[55] S. Tappertzhofen, R. Waser, and I. Valov, ChemElectroChem **1**, 1287 (2014).

[56] N. Knorr et al., J. Appl. Phys. **114**, 124510 (2013).

[57] F. Messerschmitt, M. Kubicek, and J. L. M. Rupp, Advanced Functional Materials **25**, 5117 (2015).

[58] I. Valov et al., Nature Communications **4**, 1771 (2013).

[59] Klaus J. Vetter, *Electrochemical kinetics*, Springer Verlag, 1961.

[60] S. Tappertzhofen, M. Hempel, I. Valov, and R. Waser, Mater. Res. Soc. Symp. Proc. **1330** (2011).

[61] D. Ielmini, C. Cagli, F. Nardi, and Y. Zhang, J. Phys. D Appl. Phys. **46**, 74006/1 (2013).

[62] D. Ielmini, F. Nardi, and S. Balatti, IEEE Trans. Electron Devices **59**, 2049 (2012).

[63] S. Yu and H.-S. Wong, IEEE Trans. Electron Devices **58**, 1352 (2011).

[64] S. Larentis et al., IEEE Trans. Electron Devices **59**, 2468 (2012).

[65] P. R. Mickel et al., Appl. Phys. Lett. **102**, 223502 (2013).

[66] I. Valov and G. Staikov, J. Solid State Electrochem. **17**, 365 (2013).

[67] T. Tsuruoka, K. Terabe, T. Hasegawa, and M. Aono, Nanotechnology **22**, 254013 (2011).

[68] T. Tsuruoka, K. Terabe, T. Hasegawa, and M. Aono, Nanotechnology **21**, 425205/1 (2010).

[69] R. Soni et al., J. Appl. Phys. **110**, 54509/1 (2011).

[70] A. Nayak et al., Nanotechnology **22**, 235201/1 (2011).

[71] A. Nayak et al., J. Phys. Chem. Lett. **1**, 604 (2010).

[72] F. Nardi et al., IEEE Trans. Electron Devices **59**, 2461 (2012).

[73] D. Kamalanathan, U. Russo, D. Ielmini, and M. N. Kozicki, IEEE Electron Device Lett. **30**, 553 (2009).

[74] J. van den Hurk, S. Menzel, R. Waser, and I. Valov, J. Phys. Chem. C **119**, 18678 (2015).

[75] S. Tappertzhofen et al., IEEE Electron Device Lett. **35**, 208 (2014).

[76] Z. Xu et al., {ACS} nano **4**, 2515 (2010).

[77] Q. Liu et al., Adv. Mater. **24**, 1844 (2012).

[78] Y. Yang et al., Nature Communications **3**, 732 (2012).

[79] Y. Yang et al., Nat. Commun. **5**, 4232/1 (2014).

[80] U. Celano et al., Nano Letters **14**, 2401 (2014).

[81] U. Celano et al., Electron Devices Meeting (IEDM), 2013 IEEE International, 21.6.1 (2013).

[82] R. Waser, R. Dittmann, G. Staikov, and K. Szot, Adv. Mater. **21**, 2632 (2009).

[83] S. Gao et al., The Journal of Physical Chemistry C **116**, 17955 (2012).

[84] Shanshan Peng et al., Appl. Phys. Lett. **100**, 072101 (2012).

[85] Q. Liu et al., Advanced Materials (deerfield Beach, Fla.) **25**, 165 (2013).

[86] S. Gao et al., J. Phys. Chem. C **117**, 11881 (2013).

[87] I. Valov and R. Waser, Advanced Materials **25**, 162 (2013).

[88] I. Valov and R. Waser, J. Phys. Chem. C **117**, 11878 (2013).

[89] S. Tappertzhofen, I. Valov, and R. Waser, Nanotechnology **23**, 145703 (2012).

[90] S. Long et al., Appl. Phys. Lett. **102**, 183505 (2013).

[91] C. Chen et al., Appl. Phys. Lett. **103**, 043510 (2013).

[92] T. Tsuruoka et al., Nanotechnology **23**, 435705 (2012).

[93] J. R. Jameson et al., IEEE Electron Device Lett. **33**, 257 (2012).

[94] S. Menzel, U. Böttger, and R. Waser, J. Appl. Phys. **111**, 014501 (2012).

D3 Valence change in Nanoionic Oxide Cells

Regina Dittmann
Forschungszentrum Jülich (PGI-7), Germany

Contents

1 Introduction **2**

2 Redox reactions and valence changes **3**

3 Common processes during VCM switching **4**

4 Forming of resistively switching cells **5**
4.1 Ionic processes 5
4.2 Location of the forming process 6
4.3 Structural changes during the electroforming process 9
4.4 Electronic processes 9
4.5 Forming free devices 11

5 Switching mechanism **11**
5.1 Experimental evidence for valence changes during switching 11
5.2 Modelling of the switching process 13
5.3 Toggling of the switching polarity 16
5.4 Interface and surface reactions 18
5.5 Cationic motion 19

6 Stability of the resistive states **20**

7 Summary **22**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

One possible approach to realize devices with memristive behaviour are metal-oxide-metal structures, which show hysteretic current-voltage (I-V) curves as shown in figure 1. Since most oxide materials used for this purpose are insulators, the devices are highly insulating in the virgin state and have to be transferred to a switchable state by applying an electrical stimulus. For many systems this, so called forming process, takes place during the first voltage sweep as depicted in grey in figure 1, but it can also be performed by either applying a DC voltage or voltage pulses prior to the I-V measurements. The subsequently recorded IV-curves follow the red curve depicted in figure 1 in a reversible way, which can be used for non-volatile data storage. Considering that a certain voltage is needed to induce resistance changes, low current sweeps can be employed to read-out the two resistive states (logic 1 or 0) of the device. However, the fields of application go beyond non-volatile memories and cover novel logic circuits as well as neuromorphic computing as will be shown in contribution E4 in great detail.

Based on the current knowledge, these reversible changes in the device resistivity can be ascribed to electrically induced redox-processes taking place in the oxide and/or at the oxide electrode interface. In most cases, the redox-process in the metal-oxide goes along with a change in the valence state of the metal ion. Therefore, this type of switching mechanisms is also called valence change mechanisms (VCM)[1].

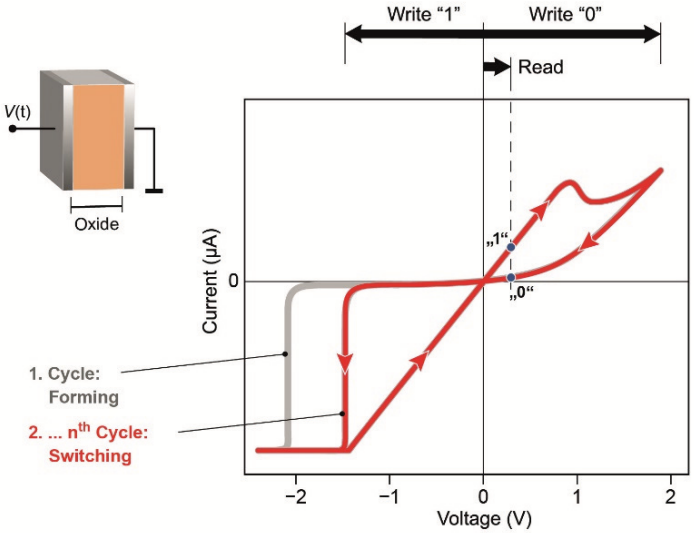


Fig. 1: *I-V curve of a resistive switching metal-oxide-metal structure.*

In this contribution, we will present the current knowledge about the microscopic mechanisms taking place during both, electroforming and resistive switching in VCM cells. Since microscopic changes during electrically biasing do generally not take place over the whole oxide device but within nanoscale filaments or at the vicinity interface, it is a challenging task to gain experimental evidence of the underlying redox-processes. We select a few model material systems where explicit experimental investigations of changes in the atomic and electronic structure during electroforming and/or switching are available. One of the main model systems chosen in this contribution are epitaxial $SrTiO_3$ thin film devices, since the defect chemistry of this

prototypic perovskite system (see figure 2(a)) and the ionic motion in this mixed ionic-electronic conductor are well elaborated and have been discussed in great detail in contributions A3-A6. Furthermore, the crystallinity in this system facilitates the observation of structural changes, which are much more difficult to separate in polycrystalline and amorphous thin films with limited long-range order. In particular, crystalline model systems offer the possibility to assign certain device properties to the presence of extended defects. However, one should keep in mind that this model system is of minor relevance for complex circuits integrated with CMOS technology as discussed in the contribution E1-E4.

We will relate the experimental findings to current approaches to simulate static I-V curves as shown in figure 1. The related switching kinetics, however, will be discussed within contribution D4 in detail.

2 Redox reactions and valence changes

A redox process denotes a coupled reduction and oxidation reaction, i. e. an electron transfer reaction where the reduction is the uptake of electrons and the oxidation is the release of electrons by atoms (see more details in contribution A6). Often, the atoms involved completely lose or win one or more electrons in their outermost electron shell, i. e. there is an integer change in their valence. This is clearly the case for localized electrons, for example, in metal ions in aqueous solutions or in the ions of an ionic solid. The picture of integer valence changes still holds when covalent bond contributions appear. However, as soon as electron delocalization by metallic bond contributions have to be taken into account, such as in non-stoichiometric compounds with delocalized electrons, redox reactions may act on a large ensemble of electrons and the result of the redox process may be a minor change in the Fermi energy level and/or the electronic band structure. Formal valence changes (per atom) would be fractional numbers in this case.

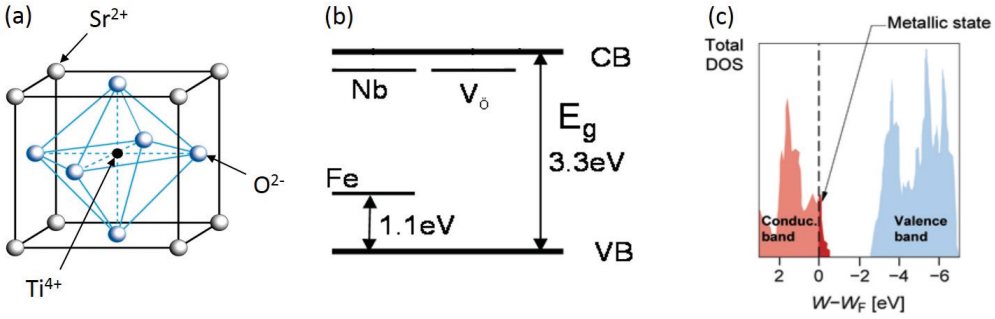


Fig. 2: *(a) Sketch of the perovskite crystal structure of $SrTiO_3$; (b) Energy levels of different point defects (Fe, Nb substitutional defects, oxygen vacancies) in $SrTiO_3$; (c) Density of states of $SrTiO_3$ with oxygen vacancies [2]*

In the VCM memories described in this contribution, the redox reactions must be coupled to the transport of ions (typically over lengths of few nanometers only) – this is why the mechanisms can also be regarded as a nanoionic redox processes.

The reduction of a metal oxide might go along with the excorporation of oxygen and the formation of oxygen vacancies $V_O^{\bullet\bullet}$ according to the following reaction:

$$O_O \rightarrow \frac{1}{2}O_2(g) + V_O^{\bullet\bullet} + 2e' \quad (1)$$

Figure 2 shows the defect levels for different kinds of point defects in the model system $SrTiO_3$. It can be seen that the defect levels of doubly ionized oxygen vacancies $V_O^{\bullet\bullet}$ are situated just below the conduction band and therefore act as donor-type defects. As a result of the formation of oxygen vacancies, the Fermi-level of reduced $SrTiO_3$ is shifted into the conduction band (figure 2(c)) which in a simple ionic model consists of Ti 3d states. As a result, the formation of oxygen vacancies goes along with a change of the formal Ti valence state from 4+ to 3+.

Therefore, a change of the oxygen vacancy concentration in $SrTiO_3$ induced by electrical biasing should go along with a population of the Ti3d band and the formation of Ti^{3+} which might be detected by spectroscopic techniques.

3 Common processes during VCM switching

Figure 3 shows a VCM metal-insulator-metal (MIM) cell (e.g. $SrTiO_3$) under current load. All conceivable processes taking place during electroforming and resistive switching of VCM cells are sketched schematically. In the simplest case, the metal electrodes M and M´ only carry electronic currents while the oxide may carry electronic and ionic currents. Common to all redox-based ReRAM is an ion current in the metal oxide and a reduction and/or an oxidation process in the cell. Given the current direction in Figure 3, the ionic current may consist of anions O^{2-} and of metal cations moving to the left and to the right, respectively. The relative current contributions strongly depend on the specific material combination and the operation conditions. Joule heating will typically occur in the interior of the oxide layer and/or close to a contact. The ionic partial current in the oxide leads to electrochemical reactions, oxidation at the anode and reduction at the cathode. The specific electrochemical interface reaction is determined by the specific material combination of VCM cell.

For noble metals, the ionic current is assumed to be blocked at the electrode interfaces. This leads to a so-called concentration polarization, i. e. an accumulation of the mobile ions near one electrode and a depletion near the other. Except in the (typically very narrow) space charge regions, this process is compensated by local redox reactions, i. e. a change in the average valence of the metal ions.

For non-noble metal electrodes, the metal might be oxidized during the anodic oxidation instead of the release of oxygen. However, the oxidation of the metal electrode might already take place during the deposition of the stack which goes along with the formation of oxygen vacancies in the metal oxide layer. Since this type of resistive switching cell has a variety of advantages, which will be explained in detail later, it is very common to intentionally grow layer stacks consisting of a substoichiometric metal oxide and the corresponding stoichiometric oxide. Another common approach is to use bilayer stacks of different metal oxides such as TiO_2/HfO_2 or Ta_2O_5/Al_2O_3.

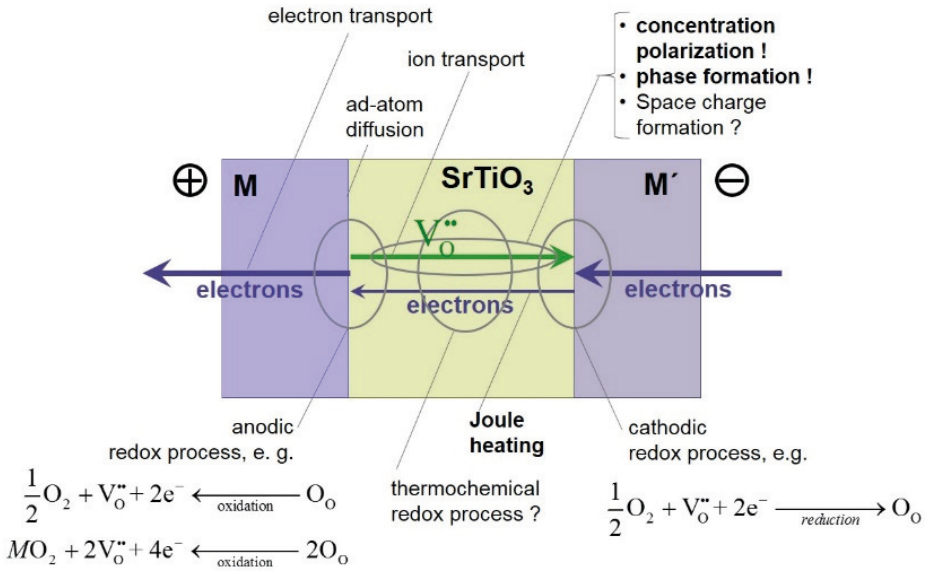


Fig. 3: *Overview of all processes which may be relevant in metal oxide (e.g. $SrTiO_3$) VCM cells. M and M´ denote the electrodes. Modified from [3].*

4 Forming of resistively switching cells

The grey curve in figure 1 shows the forming step of a resistive switching cell which turns the metal oxide insulator into a reversibly switchable material with increased electronic conductivity. This process is often also regarded as a soft-breakdown process, in the sense that the process can at least be partly reversed by an electrical stimulus. Complex processes might take place during breakdown of metal oxides, comprising electronic as well as ionic processes, often mediated by Joule heating and thermal runaways. The microscopic processes taking place during electroforming might vary strongly with the material system and the details of electrical biasing. In the following we will present a few examples of how electroforming process might be described and what kind of microscopic changes have been observed experimentally.

4.1 Ionic processes

One possibility to explain electroforming is to assign it solely to ionic motion and the resulting electrode reactions described in the previous section. We assume a cell with an chemically inert active electrode and an ohmic counter electrode with a high oxygen affinity (e. g. a cell $Pt/TiO_2/Ti$). Electronically, this MIM cell corresponds to a Schottky diode.

A positive forming voltage will lead to an electroforming into the high resistive state (HRS). The Schottky diode is biased in forward direction. The anodic oxidation reaction will release O_2 gas and may lead to entrapped gas bubbles underneath the Pt electrode as shown in figure 4.

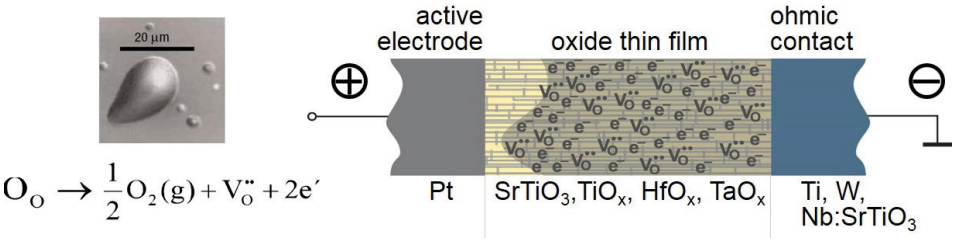


Fig. 4: *Sketch of the electroforming process based on anodic oxidation and the propagation of a virtual cathode formed by the front of oxygen vacancies. The left picture shows bubbles formed in the Pt electrode by the formation of oxygen gas at the anode [4].*

Oxygen vacancies $V_O^{\cdot\cdot}$ are injected into the oxide and drift rapidly towards the cathode due to the high electric field. If there is no redox reaction at the cathode interface, the $V_O^{\cdot\cdot}$ start to accumulate near the cathode, which is compensated by electrons. Along with the further accumulation of $V_O^{\cdot\cdot}$, the *n*-conducting cathodic region, sometimes called *virtual cathode*, propagates towards the anode. When only a relatively small potential barrier (disc) remains, the resistance of the cell decreases significantly (usually limited by a compliance current) and the electroforming into the HRS is completed. This process can also be used if one starts from a symmetric cell of high work function electrodes, e. g. Pt/TiO_2/Pt, because a virtual cathode is formed during the reduction process. This type of forming process has been numerically simulated for $SrTiO_3$ cells with two blocking electrodes by a drift-diffusion simulation of the time evolution of the oxygen vacancy distribution [5].

A negative forming voltage will lead to an electroforming into the low resistive state (LRS). The anodic oxidation reaction will lead to an oxidation of the Ti electrode in this case. Oxygen vacancies $V_O^{\cdot\cdot}$ are formed and drift towards the Pt cathode. The difference to the process for forming into the HRS is, that the anode is a low work-function metal so that no effective barrier remains at the end of the process. In addition, Ti consumes oxygen ions also purely chemically, to that no fully oxidized TiO_2 layer is left.

With respect to the direction of the movement of the front of oxygen vacancies one has to consider the relative kinetic limitations of oxygen excorporation and oxygen drift [6]. As a result, one can show that irrespective of the applied polarity the front of oxygen vacancies always moves from the reactive metal electrode where the supply with oxygen vacancies is strongly enhanced with respect to an inert electrode where oxygen gas has to be excorporated.

4.2 Location of the forming process

In order to distinguish if the electroforming procedure results in the formation of a homogeneous front of oxygen vacancies as sketched in figure 4 or is restricted to a filament as sketched in figure 5, the Pt top electrode of single crystalline $SrTiO_3$ thin film devices was removed by a dedicated delamination process after the forming procedure [7]. Subsequently, the remaining electrode interface was investigated by conductive tip atomic force microscopy (LC-AFM).

The topography as well as the conductivity at the position of the former electrode can be seen in figure 6. These investigations show that most of the cell area is unaffected by the forming procedure, but significant changes of the morphology can be observed in one corner of the

former electrode (figure 6(a)-(b)). Cross-sectional transmission electron microscopy (TEM) investigations showed that a high density of extended defects is formed in this region [8]. However, it is important to note that only this disturbed region shows a significant current level, whereas the remaining area persists insulating during the forming procedure (figure 6 (c)-(d)). Therefore, on can conclude that electroforming results in locally restricted changes of the thin film microstructure and its electronic conductivity rather than in homogeneous changes.

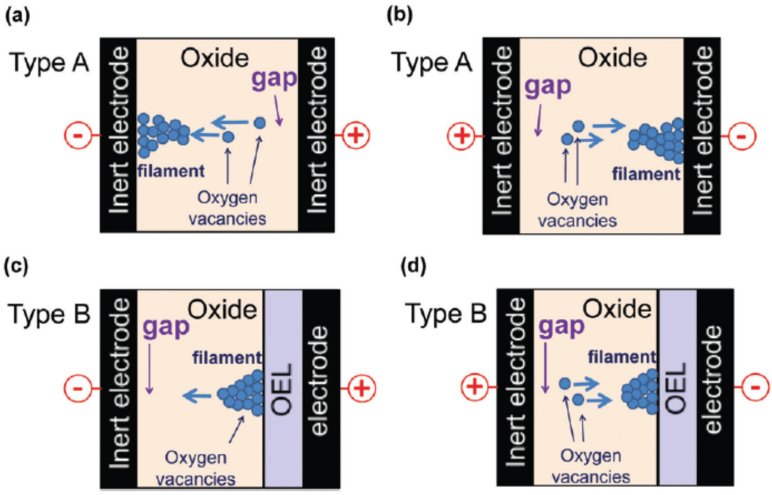


Fig. 5: *Schematic illustration of the growth of a conductive filament during electroforming for different electrodes (inert, oxidizable). [6]*

Similar Fe-doped $SrTiO_3$ cells have been investigated by micro-focused X-ray absorption spectroscopy, which in the fluorescence detection mode provides bulk information about chemical changes in the electroformed cells. For this method, the Fe doping atoms serve as tracer elements for the redox-process taking place in the device during the electroforming procedure. Figure 7(a) shows the Fe K-edge spectra recorded on different positions of an electroformed device, on a reference films as well as on a device where a complete breakdown was induced during the electrical treatments. Clear changes of the pre-edge features are visible which can be attributed a different amount of oxygen vacancies in the first coordination shell of the Fe dopants [9]. In particular, the oxygen vacancy concentration has the lowest value for the reference Fe-doped $SrTiO_3$ thin film and the highest concentration of the breakdown spectrum. Based on the pre-edge intensity at 7122eV, we determined an oxygen vacancy map of the device which is depicted in figure 7(b). It shows clear evidence for the creation of an oxygen vacancy rich filament during electroforming. Furthermore, it is important to note that the oxygen vacancy concentration beneath the electrode (see blue spectrum in figure 7(a)) is significantly higher than in the thin films reference (see black spectrum in figure 7(a)). We therefore conclude that electroforming result in both, a net increase of the oxygen vacancy concentration underneath the whole electrode as well as the formation of a vacancy rich filament which might be formed at a later stage of the electroforming procedure as sketched in figure 7(c). A broad front of oxygen vacancies might move in the early stage of electrical biasing as proposed in figure 4. However, at certain defective positions of the film, the oxygen vacancy formation or their

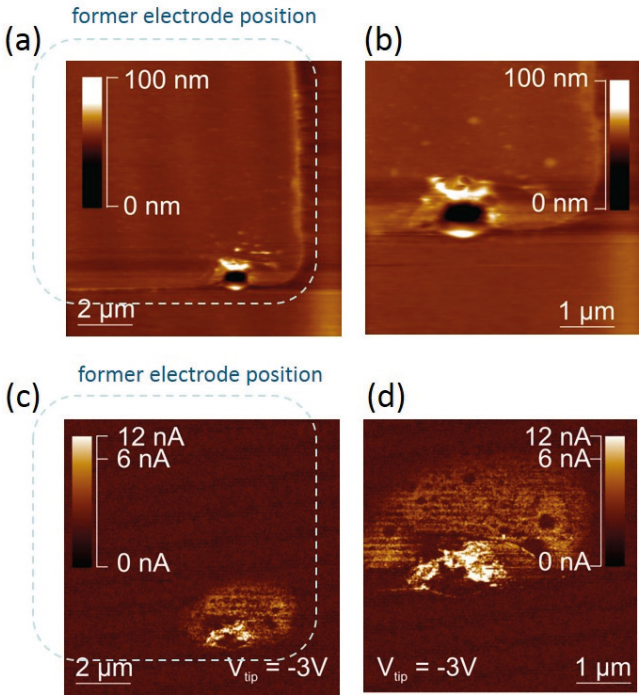


Fig. 6:*(a)-(b) AFM topography scan of a formed Pt/Fe-doped $SrTiO_3$/Nb-doped $SrTiO_3$ cell after Pt electrode delamination; (c)-(d) corresponding conductivity scans [8].*

movement might be facilitated, resulting in the formation of spikes in the moving oxygen vacancy front. As soon as one of the spikes reaches the electrode vicinity, a self-accelerating process take place according to the flow of electronic currents and the resulting Joule heating. This results in the formation of a strong filament and a disruption of the vacancy formation process in the remaining area.

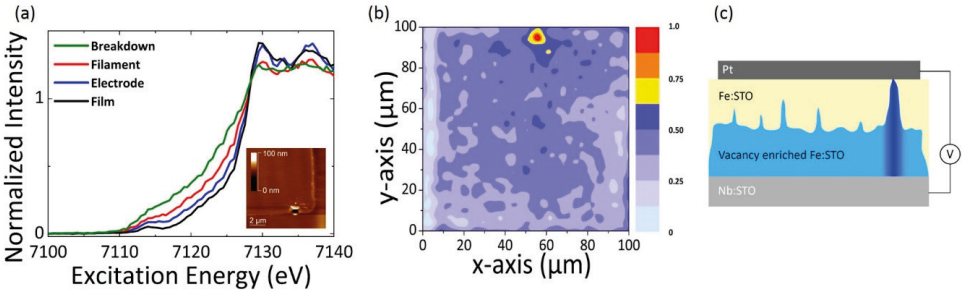


Fig. 7: *(a) Fe K-edge of a Fe-doped $SrTiO_3$ cell (fluorescence detection mode) measured at different positions; Inset: AFM scan shown in figure 6 ;(b) Oxygen vacancy map extracted determined from the pre-edge intensity at 7122eV; (c) Sketch of the forming process based on the Fe-K edge data shown in (a); extracted from [9]*

It is important to note that it has been shown by in-situ TEM analysis that electroforming in $SrTiO_3$ single crystals takes also place if no dislocations or other extended defects are present in the detected area [10].

4.3 Structural changes during the electroforming process

Although simulations of the electroforming processes based on the drift-diffusion of point defect oxygen vacancies are quite successful to describe the experimental data to a certain extent, there exist numerous reports that electroforming goes along with a considerable change of the structure. In TiO_2 thin films and single crystals, the solubility of point defect oxygen is quite low and oxygen vacancies tend to order and form Wadsley type of defects which are nucleation points for the formation of Magnéli phases Ti_nO_{2n-1} during electroreduction [11]. In situ TEM analysis could prove the reversible formation and dissociation of Wadsley defects with electrical biasing (see figure 8), which is, however, not directly correlated with the resistive switching phenomena in the devices. In particular, the movement of Wadsley defects takes place a voltages which correspond rather with the read-voltages than with the write voltages and have to be considered as possible origin of a read disturbance [12].

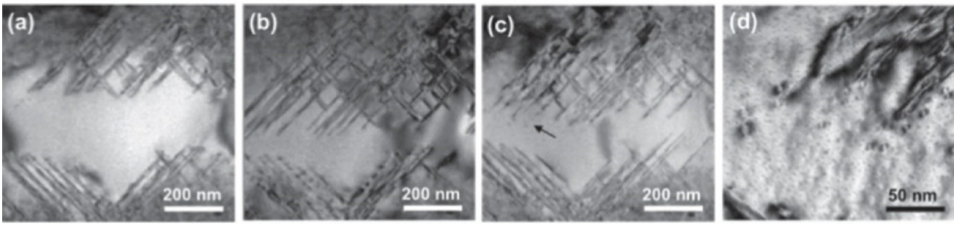


Fig. 8: *In situ TEM analysis of resistively switching TiO_2 single crystals [12]. A reversible movement of Wadsley planar faults are visible. Between (a) and (b) a voltage of -1.35V has been applied; Between (c) and (d) a voltage of +1.15V has been applied.*

Besides these reversible changes, electroforming might also lead to irreversible changes such as the creation of dislocations [10] or a phase separation of $SrTiO_3$ into a Sr deficient phase and SrO [8], [13]. Most of the observed phase formation processes have in common that a significant amount of Joule-heating is required in order to obtain a sufficient mobility of both oxygen ions as well as cation ions.

4.4 Electronic processes

Although strong experimental evidence exists that electroforming is connected with ionic movement as discussed previously, detailed studies of the early stage of electroforming have shown that electronics processes play a dominant role in initiating the forming procedure [14].

Figure 9 shows the voltage across a $Pt/Ta_2O_5/Pt$ device as a function of time. Each curve corresponds to the trace during a single voltage pulse applied to the same device. The gradual decrease of voltage in the beginning is associated with the decrease of the resistance with increasing Joule-heating. The rapid drop between 45 and 55ns corresponds to the beginning of the electroforming process. No permanent change of the device takes place after the pulses (1)-(3). Therefore, the initial part of the sharp resistance drop is volatile and has to be electronic

in nature [14]. It is important to note that the current flow induced by this electronic instability is initially uniform and can spontaneously and reversibly constrict to a localized filament. For pulse (4), the resistance drop is permanent and the electroforming process can be regarded as completed. One can conclude from these experiments that electroforming is initiated through purely electronic and reversible events, to be followed by oxygen vacancy movement or structural changes.

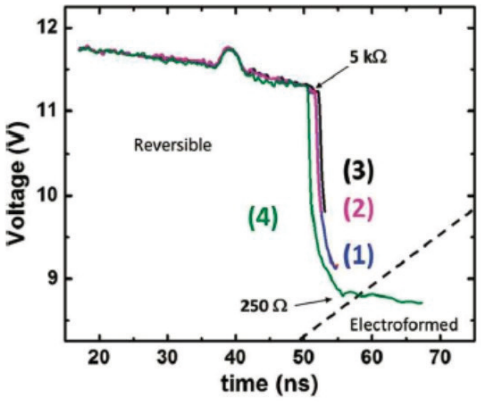


Fig. 9: *Pulsed electroforming experiments: The different voltage-time curves are determined for voltage pulses with different pulse length [14].*

Another scenario which hints on the initiation of the electroforming process by electronic effects, is the appearance of fractional, dendrite-like structures in Ta_2O_5 devices shown in figure 10 (a). Since the formation of this structure is strongly enhanced in the presence of an interface adsorbate layer, it is concluded that these structures are induced by an avalanche discharge between the top electrode and the Ta_2O_{5-x} layer as sketched in figure 10(b). The entire dendrite region becomes permanently conductive and shows a significant valence change from Ta^{5+} to Ta^{4+} [15]. Therefore, one can conclude that the avalanche discharge induces local heating which subsequently results in a reduction of Ta_2O_{5-x} in the whole dendrite area (figure 19 (c)) and a decrease in the cell resistance.

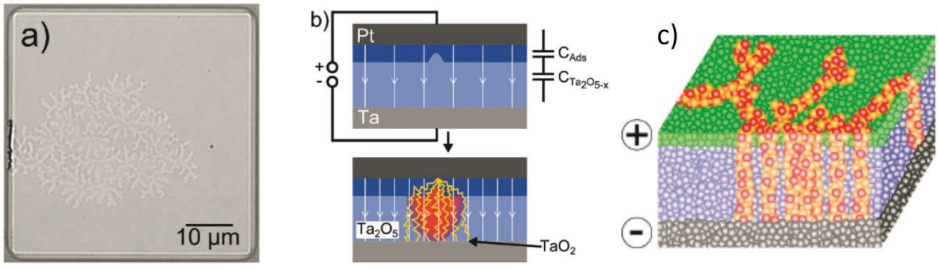


Fig. 10: *a) Optical microscope image of the Pt top electrode after electroforming. b) Scheme of the MIM structure with adsorbate layer in between the Pt/Ta_2O_{5-x} interface. The avalanche discharge preferentially takes place at the impurities. c) Schematic profile of the dendrite-like structure with an interfacial layer (consisting of H_2O and hydrocarbons) (green circles), Ta_2O_{5-x} (blue circles), Ta bottom electrode (grey circle) and the conductive path (orange/yellow circles depicting oxygen vacancies and Ta^{4+} ions) developed after forming with a positive voltage at the platinum top electrode.[15]*

4.5 Forming free devices

Tuning of the oxygen vacancy concentration in HfO_{2-x} thin films by the deposition conditions showed that the forming voltage systematically decreases with increasing oxygen vacancy concentration [16]. Since an increased oxygen vacancy concentration goes along with a higher electronic conductivity, Joule heating and temperature-accelerated ionic motion becomes significant at lower voltages. If the forming voltage approaches the switching voltage, the devices can be regarded as forming-free.

As mentioned previously, in cell stacks with oxidizable electrodes, a redox-process with the metal electrodes takes place and results in an increase of the oxygen vacancy concentration in the oxide layer. It has been shown for Nb:STO/$SrTiO_3$/Ti devices that the degree of oxidation of the Ti electrode depends on the relative thickness of STO to the Ti cap layer [17]. If the Ti thickness exceeds a certain value, the devices become forming free. This effect has been confirmed for a large variety of layer stack combinations such as and HfO_2/Hf [18]. Figure 11 shows a linear dependence of the forming voltage on the HfO_2 thickness of HfO_2/Hf cells. As can be seen from the set statistics in the inset of figure 11, a stack with 2nm thick HfO_2 can be regarded as forming-free.

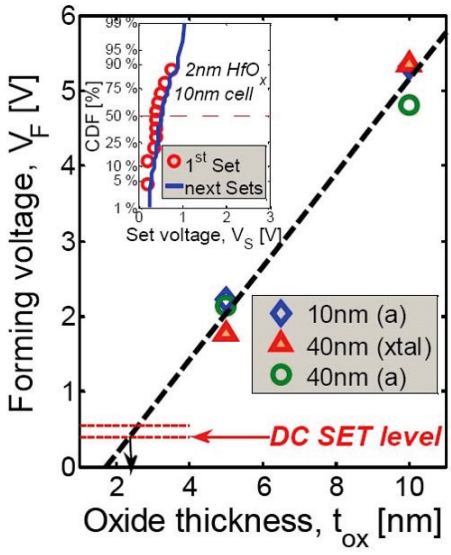


Fig. 11: *Dependence of the forming voltage on the oxide thickness. Inset: Weibull-plot for the SET process for the 1st and the next set events. The cell can be regarded as forming-free [18] .*

5 Switching mechanism

5.1 Experimental evidence for valence changes during switching

As mentioned in the previous chapter, the forming procedure results in the creation of a conducting filament. Therefore, it is likely to assume that the filament is interrupted during the RESET process and restored during the SET process. In order to clarify the microscopic origin of the

switching process in $SrTiO_3$/Pt devices, spectromicroscopic investigations have been performed on devices in the LRS and in the HRS, respectively, after delaminating the top electrode [13]. The delamination of the top electrode is necessary since the employed X-ray absorption spectroscopy has an insufficient probing depth to detect chemical changes beneath the Pt electrode layer. Figure 12 shows the Ti L-edge in different regions of a device in the HRS and in LRS, recorded in a photoemission electron microscope (PEEM). For the sample in the LRS, a clear filament region can be identified where the Ti K-edge shows strong deviations from the typical Ti^{4+} spectra recorded on the remaining device area. A comparison of the spectrum with reference data from the literature [19] clearly proves the presence of Ti^{3+} in this filament region. Based on a principle component analysis, we determined Ti^{3+} maps for the two devices in the different resistive states which can be seen in the left of figure 12. These results clearly prove that resistively switching of the Pt/$SrTiO_3$ devices can be assigned to a valence change on the Ti site.

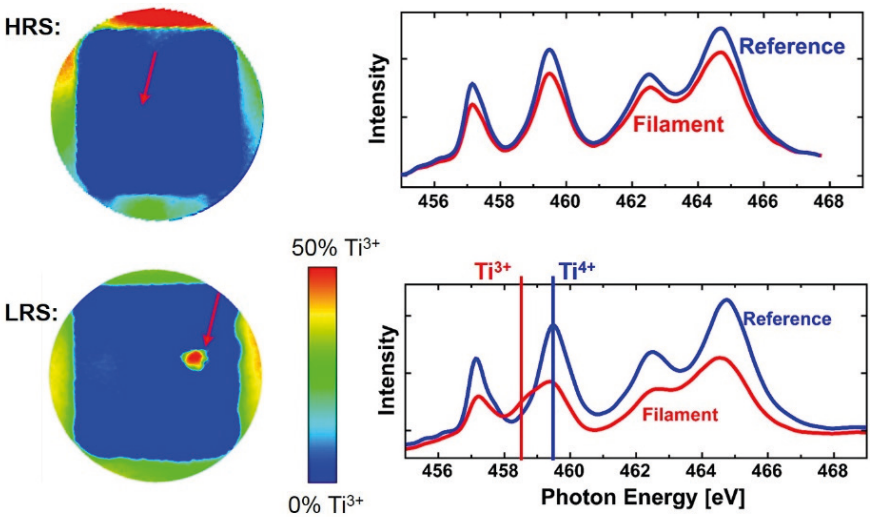


Fig. 12: *PEEM analysis of $SrTiO_3$ cells in HRS and LRS. On the right are depicted the Ti L-edge spectra within the filament and outside. The left shows the false colour maps of the Ti^{3+} content determined by principle component analysis [13].*

The same experimental approach has been used for the analysis of the Ta/Ta_2O_{5-x} /Pt devices which exhibit a dendrite-like pattern after electroforming as shown in figure 10. Figure 13 shows the Ta 4f core level X-ray photoelectron spectra for the dendrite region in LRS (c) and HRS (d), respectively. The spectra have to be fitted by a Ta^{5+} (blue spectrum) as well as a Ta^{4+} (red spectrum) contribution. It can be clearly seen that the Ta^{4+} contribution is strongly increased in the LRS state. Figure 13(a) and (b) shows the lateral distribution of the Ta valence state within the device area.

Whereas the device in the HRS shows almost no Ti^{4+} contribution, the sample in the LRS shows a significant Ta4+ contribution in nearly the whole dendrite area. Based on the analysis one can conclude that resistive switching in these devices is based on a valence change in the Ta_2O_5 which results in the formation of the metastable TaO_2 phase within the dendrite region [15]. As the inelastic mean free path for photoelectrons at a kinetic energy of 1460 eV is approximately

2.3 nm [21], we can estimate that the depth underneath the top Pt electrode (called disc) where a valence change takes place during resistive switching is greater than or equal to 2.5 nm.

While PEEM analysis of devices is limited to detect filaments exceeding the 20nm scale, cross sectional high resolution annular dark field (HAADF) investigation in a scanning TEM (STEM) provide atomic resolution. Figure 13(e) and (f) show HAADF-STEM image of a Pt/TaO_{2-x}/Ta_2O_{5-x}/SiO_2/Pt device stack in the HRS and LRS, respectively. During electroforming, the Ta_2O_5 has moved into the SiO_2 interlayer and has formed a network of 1nm long filaments which are marked with yellow arrows. During resistive switching, a strong change of the contrast in the filament region is observed which can be attributed to a change of the Ta_2O_{5-x} to lower valence states [20]. These atomic scale investigation show that valence changes might not occur at a single compact filament, but might be fragmented into multiple filaments at the active device electrode.

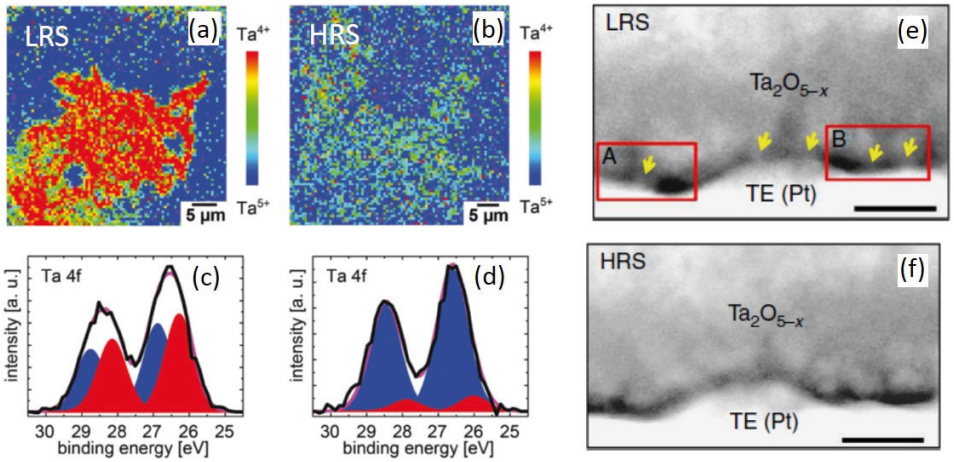


Fig. 13: *(a)-(d) PEEM analysis of a Ta/Ta_2O_{5-x} /Pt device shown in figure 10(a) after Pt delamination. False color maps showing the distribution of Ta^{4+} (yellow) and Ta^{5+} (blue) for the LRS a) and HRS b). Ta 4 f core level spectra from the dendrite-like structure (red curve) and the reference region (blue curve) for an LRS c) and HRS c) cell. The spectra were fitted with two components, one for Ta_2O_{5-x} (blue) and one for the low binding energy component TaO_2 (red), the total fit is presented by the pink curve[15]. (e)-(f): In situ HAADF-STEM analysis of a Pt/TaO_{2-x}/Ta_2O_{5-x}/SiO_2/Pt device on LRS (e) and HRS (f), scale bars: 5nm[20].*

5.2 Modelling of the switching process

Based on the experimental findings in the different systems described above it is reasonable to assume that resistive switching goes along with a valence change within the filaments created during the electroforming procedure. As will be shown in detail in contribution D4, one has to assume that this valence change is restricted to a nanoscale region in the vicinity of the active electrode in order to realize a sufficient fast movement of oxygen vacancies to enable fast switching events in the ns scale.

Figure 14(a) illustrates a simplified view of the switching process based on these considerations. A typical MIM systems consists of an active interface (active electrode, AE), at which the switching takes places, a mixed ionic-electronic conducting (MIEC) I-layer, and an ohmic counter electrode (OE). A typical I-V characteristic is shown in figure 14(a), together with sketches of the switching mechanism. In the HRS (Figure 14(a) A), the filament consists of the n-conducting MIEC oxide (called plug) and a potential barrier (called disc) in front of the left electrode. Upon application of a negative voltage, oxygen vacancies from the plug part of the filament are attracted into the barrier (Figure 14(a) B), which results in a significant decrease of the barrier height due to a local reduction process, which turns the cell into the LRS (Figure 14(a) C). For the RESET, a positive voltage is applied to the active interface which repels the oxygen vacancies (Figure 14(a) D), leading to a local re-oxidation, and turns the cell into the HRS again.

The band diagram of the active interface in the LRS and HRS is shown schematically in Figure 14(b). Because of the small dimensions, the exact potential landscape will be determined by the local density of states (LDOS) which deviates from this macroscopically defined band picture. For a qualitative discussion of the states, however, the band picture approximation is sufficiently precise. In the OFF state, a significant energy barrier exists which originates from the fact that the disc region is fully oxidized. The temperature dependence of the resistance in the OFF state R_{OFF} typically shows a thermally activated characteristics. In our picture, it will be mainly determined by $R_{disc,OFF}(T)$ if the cells - and, hence, the conduction through the remaining area is not too large. Presumably $R_{disc,OFF}(T)$ is caused by a thermionic transmission over the barrier. An additional current contribution may be due to direct tunneling or Fowler-Nordheim tunneling. In any case, as expected for any potential barrier, $R_{disc,OFF}$ shows a strong voltage dependence, i. e. it is a highly non-linear resistor which gives rise to pronouncedly non-linear *I-V* characteristics.

During SET, oxygen ions are removed from the disc. For homogeneous systems, this is identical with an injection of oxygen vacancies which will be (at least partially) compensated by electrons. A local lattice rearrangement may take place as well, which could be described in terms of a local phase transformation. In any case, the extraction of oxygen ions will lead to a (chemical) reduction of the disc region which in turn will result in a significant decrease in the barrier width and/or height (Figure 13b). In the LRS, as a consequence, there is a pronounced increase of the thermionic conduction and of the tunneling conduction. R_{ON} is determined by the series combination $R_{disc,ON}$ and R_{plug}, while the contribution from the remaining area can be neglected. Therefore, the temperature dependence of R_{ON} might be dominated by R_{plug}. $R_{plug}(T)$ is supposed to be weak because the n-conducting oxide represents a semiconductor in the saturation regime or (more often) a degenerate semiconductor which can be described as a metal (Sec. 3.4). During RESET, oxygen ions are attracted into the disc, leading to the HRS again.

Based on the idea of the vacancy modulation of the Schottky-barrier described above, 1D numerical simulation have been performed within a drift-diffusion model of electronic-ionic transport in an n-conducting mixed ion-electronic conductor [22]. The two electrodes of the MIM structure are assumed to form ion-blocking Schottky-barriers with different barrier height. Therefore, the simulations also include the scenario of one Schottky-contact and one ohmic contact as described above. The current transport in the MIM structures occurs via electron tunneling and thermionic emission. Capturing the transition between Schottky and ohmic contact resistance upon temperature-accelerated ion migration induced by Joule heating, the model describes experimentally observed IV curves.

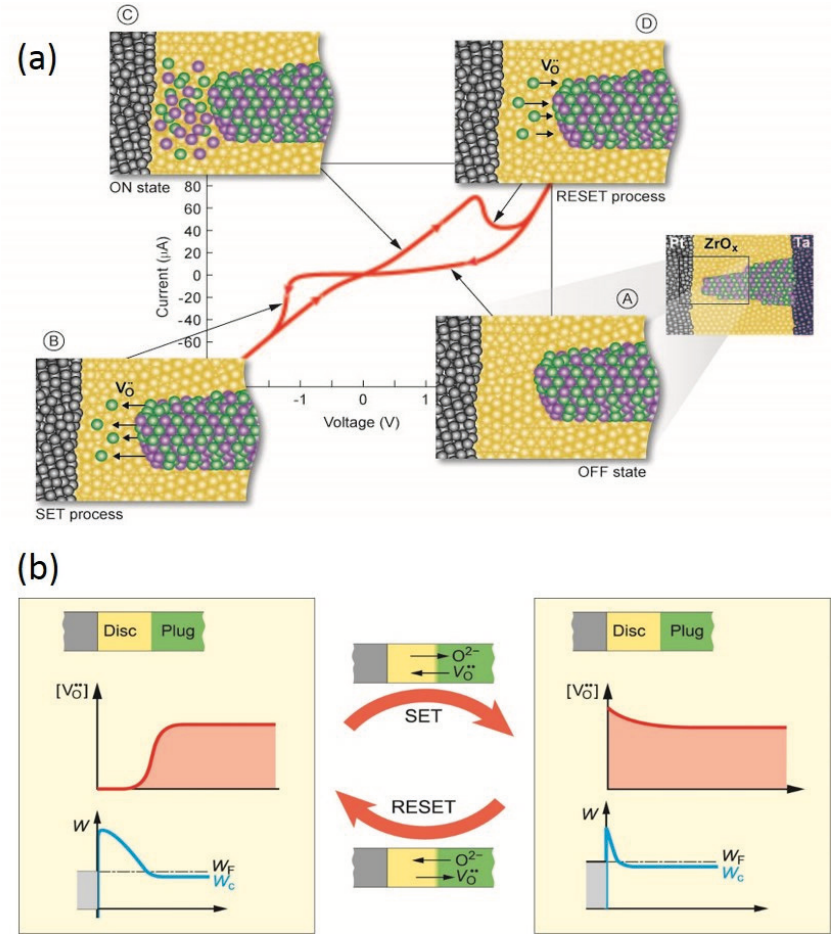


Fig. 14:*(a) Schematic I-V curve of a VCM cell together with the different stages of the switching procedure in the vicinity of the active interface. The green spheres indicate oxygen vacancies, the purple spheres indicate metal ions in a lower valence state. (A) OFF state (HRS); (B) SET process; (C) ON state (LRS); (D) RESET process. (b) Illustriation of the oxygen vacancy distribution and the band-diagram at the interface after SET and RESET.[3]*

Figure 15 shows the quasistatic I-V curve based on Schottky-barrier heights at the two interfaces of 0.7eV and 0.1 eV, respectively. Furthermore, at 7 different stages of the I-V curves, the oxygen vacancy profiles as well as the band-diagram is depicted. It can be clearly seen that the most pronounced difference between LRS and HRS is the modification of the barrier width at the interface with the higher barrier height, forming the active interface.

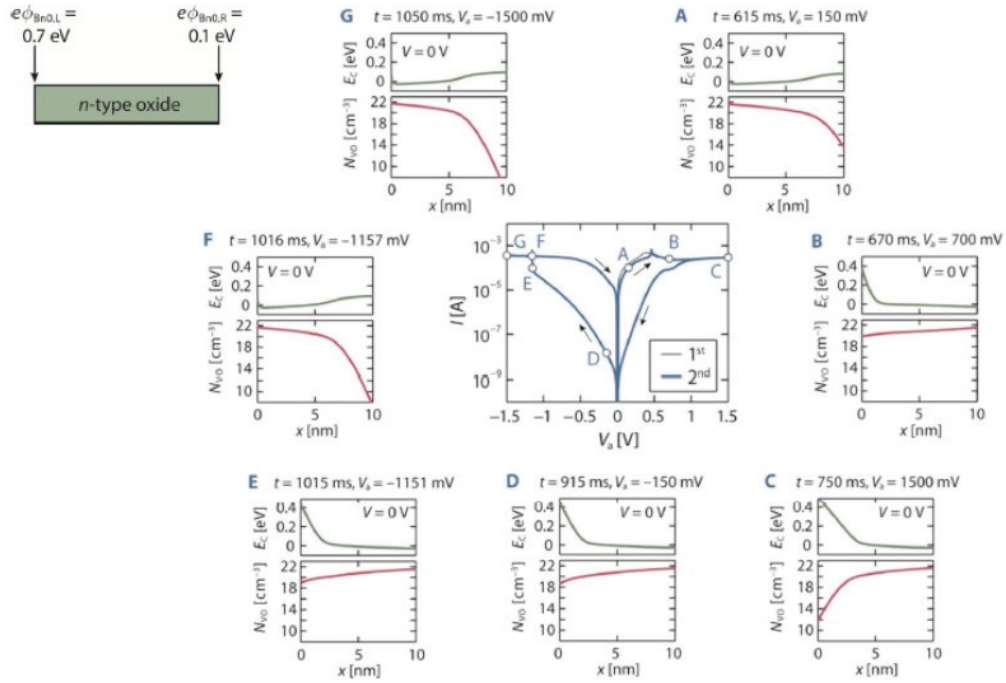


Fig. 15: *1D numerical simulation of the IV-curves by drift-diffusion of electronic-ion transport combined with asymmetric Schottky barriers at the two interfaces [22]. A-G: spatial dependence of the oxygen vacancy concentration and the band-diagram within different stages of the hysteresis curve [23]*

5.3 Toggling of the switching polarity

In the simple consideration that one single interface is the active interface and the other interface acts as starting point for the filament-shape virtual cathode, the switching polarity is well defined to be counter-eightwise (as depicted in figure 14) assuming that the active interface is biased and the other interface is grounded. However, it is often reported in the literature that the switching polarity can be turned by slight modifications of the switching stacks or that the switching polarity can be modified by the biasing procedure. Figure 16 shows the I-V curve of a forming-free Pt/Ti/TiO_2/W stack for the first and second cycle, respectively. Whereas the cell shows counter-eightwise switching polarity in the first cycle, the polarity is turned to eight-wise polarity in the second sweep. By carefully adjusting the biasing scheme, one switching polarity could be stabilized [24]. This observation implies that the active interface is changing for different biasing schemes and the related oxygen vacancy distribution at the interface. Since W and Ti are both low work function metals, slight changes in the barrier heights could result in a change of the active interface.

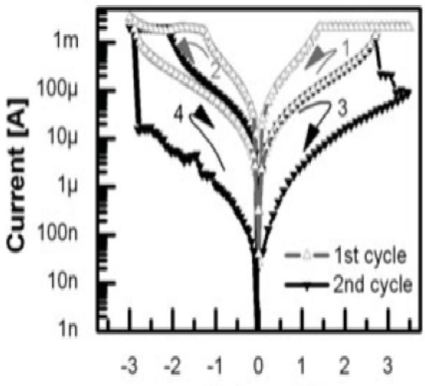


Fig. 16: *I-V curve of a Pt/Ti/TiO2/W stack. The switching polarity is counter eightweise for the first cycle and changes to eightwise in the second cycle [24].*

By changing the relative Schottky-barrier heights at the two electrodes in the drift-diffusion model [22] mentioned above one can obtain both switching polarities, namely, eightwise and counter-eightwise as well as the so called complementary CS switching when both interfaces have similar interface barriers. The case of CS switching can be regarded as intrinsically anti-serial connected resitive switches, so called complementary resistive CRS switches [25], which will be discussed in detail in contribution E4.

The transition between bipolar and CS switching has been experimentally observed for several material systems such as Pt/Ta2O5/Ta/Pt which is illustrated in Figure 17 [26]. Whereas devices with 15nm thick Ta electrodes show well pronounced counter eightwise switching (fig-

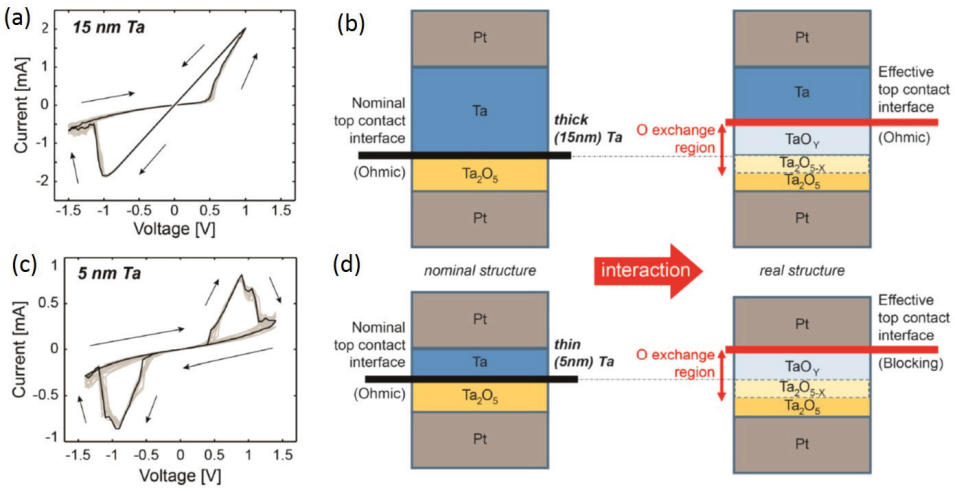


Fig. 17: *Modification of the switching mode by varying the Ta electrode thickness of Pt/Ta2O5/Ta/Pt stacks; (a) For 15nm Ta thickness the cell show bipolar counter-eightwise switching, whereas for 5nm Ta thickness CS switching is observed (c). A possible scenario for the interface configuration for (b) 15nm Ta and (d) 5nm Ta [26].*

ure 17(a)), cells with 5nm Ta exhibit CS switching after a few cycles (figure 17(c)). One possible explanation is illustrated in figure 17(b) and (c). Due to the reduced thickness of the Ta layer, it might be completely oxidized and result in a more or less symmetric stack with two Pt electrodes. Another possibility could be that Ta sucks a different density of oxygen during the interface reaction and thereby changes its workfunction for the different thicknesses [27]. As a result, the workfunction of a 5nm thin Ta electrode with increased relative density of oxygen could become similar to the Pt electrode resulting in a symmetric stack which exhibits CS switching [26].

Besides the toggling of the switching polarity by the electrode thickness in polycrystalline binary switching systems, it has been shown in crystalline $SrTiO_3$ and Sr_2TiO_4 thin films grown on Nb-doped $SrTiO_3$ that the switching polarity can be turned by the film thickness or its defect density [28]. Whereas for thick films and high defect densities, both switching polarities coexist [8], for ultrathin films and low defect density the eightwise polarity dominates [28].

5.4 Interface and surface reactions

A large variety of experimentally observed phenomena can be nicely explained by assuming a redistribution of oxygen vacancies within the layer stacks. In that case, drift-diffusion models considering oxygen blocking electrodes [22] can sufficiently describe the microscopic switching mechanism. However, electrode reactions during fabrication as well as during electrical biasing imply that they might play a role during the switching process as well. Indeed, in operando X-ray photoelectron spectroscopy on Ti/HfO_x/TiN devices gave experimental hints that the Ti electrode is directly involved in the switching process. Figure 18 shows the Ti core levels of Ti/HfO_x/TiN cells in different resistive states [29]. During electroforming, a significant decrease of the metal contribution as well as an increase in the Ti^{4+} contribution has been observed. The difference between LRS and HRS is less pronounced, however, a small shift of the spectral weight from high Ti valence states (Ti^{3+}) to low valence Ti states (Ti^{1+}) has been observed during SET operation (see figure 18 (d)). The observed eightwise switching polarity is consistent with a reversible redox-reaction between a TiO_x layer at the bottom electrode and the HfO_x switching layer, resulting in a change of the band-bending at the Ti/HfO_x interface [29].

While it is straight-forward to discuss a redox-process between oxidizable electrodes and switching oxides, a variety of experiments also hint on a redox-process taking place at noble metal electrodes such as Pt although this process can be regarded as less likely since Pt is generally regarded as oxygen blocking electrode. One possibility suggested is the formation of PtO_x [30] which, however, requires a high formation energy. Since it has been clearly proved that oxygen is excorporated during electroforming, resulting in a formation of bubbles underneath the Pt top electrode as shown in figure 4, one could also speculate of an oxygen excorporation/incorporation as possible switching mechanisms. In that case, the redox-reaction might be restricted to the three phase boundary or oxygen transport has to take place via grain boundaries or cracks in the Pt electrode. Since a large variety of publications show a dependence of the resistive state level (e.g. [31]) or the reset probability [32] on the ambient oxygen pressure, excorporation/incorporation as switching process has to be considered as possible scenario for noble metal electrodes.

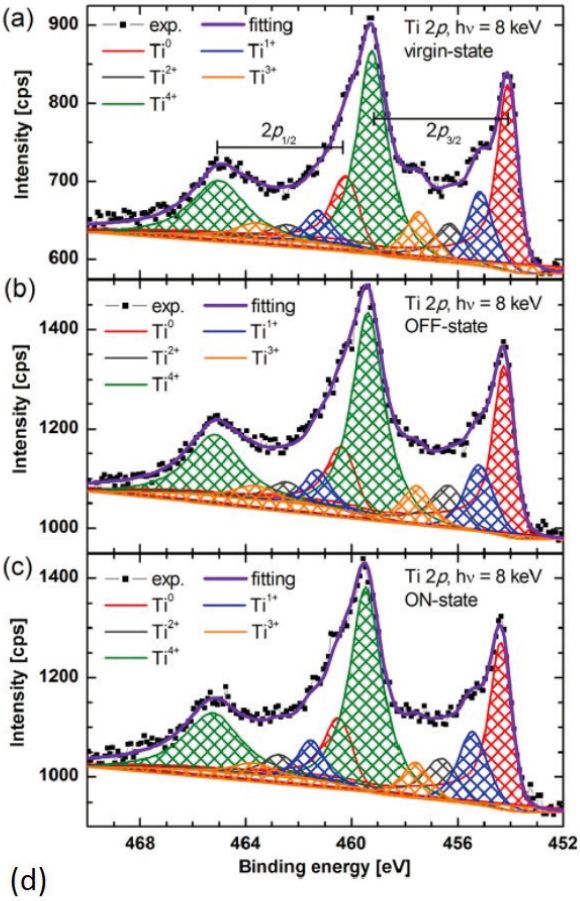


Fig. 18: *Ti2p core-level spectra measured on Ti/HfO$_x$/TiN devices with a photon energy of 8eV in different resistive states with fit curves considering 5 different Ti oxidation states.(a) virgin state, (b) OFF-state (HRS), (c) ON-state (LRS). In (d) the relative intensities of the oxidation levels in the different resistive state have been summarized [29].*

Φ ± 0.2 (%)	Ti^{0}	Ti^{1+}	Ti^{2+}	Ti^{3+}	Ti^{4+}
Virgin	37.3	9.9	10.8	7.1	34.8
OFF	34.3	**9.9**	12.8	**5.9**	37.1
ON	34.5	**10.8**	12.8	**4.9**	36.9

5.5 Cationic motion

In the previous part of this contribution, resistive switching has been described as the movements of oxygen vacancies, however, cation interstitials might be another possible candidate for a donor-type defect. For crystalline perovskite materials such as $SrTiO_3$ with a closely packed lattice, interstitial defects can generally be excluded. For binary oxides, cation interstials are well accepted as donor-type defects, however the relative mobility of oxygen vacancies and

cation interstitials is under controversial debate. It has been recently suggested that Ta interstials might become the dominant donor-type defect responsible for resistive switching in $Ta/Ta_2O_5/Pt$ cells [33], [34]. By inserting a graphene layer at the Ta/Ta_2O_5 interface, the oxidation of TaO_x is suppressed resulting in a filament growth based on Ta interstitials rather than on oxygen vacancies (see figure 19(a)). As a result, the cells show I-V curves (see figure 19(b)) which can assigned to an electrochemical metallization cell (ECM) type of switching addressed in contribution D2.

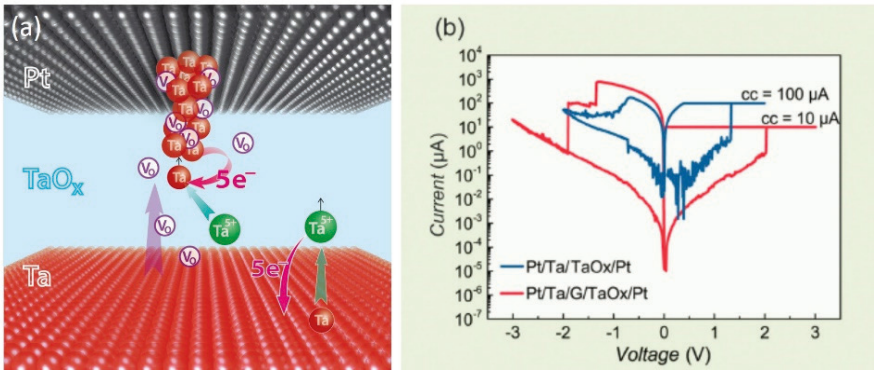


Fig. 19: *(a) Sketch of the formation of a Ta filament in $Pt/Ta_2O_5/Ta$ cells; Modification of the switching mode from VCM to ECM switching by the introduction of a graphene layer at the Ta interface [33], [34].*

6 Stability of the resistive states

Many research groups in academia and industry have already presented devices with excellent device performance. The ultimate aim regarding device performance is an endurance (write cyclability) of at least 10^7 cycles [1]. Most of all, the ultra-nonlinear switching kinetics (also called voltage-time dilemma) between extremely fast switching times ($\leq$ 10-100 ns) and long retention times (exceeding 10 years) has to be met for non-volatile memory applications.

In this contribution, we have so far only focussed on the description of the static redox-processes taking place during electrical biasing such as during electroforming and resistive switching. All aspects connected with the kinetics of the switching processes such as the movement of oxygen and the related energy barriers which have to be overcome during resistive switching will be considered in great detail in contribution D4. Furthermore, reliability issue such as device failure will be the content of E1.

Therefore, in the following we will only briefly describe some basic aspects of data retention of VCM cells. For any discussion of the stability of states and the long term reliability, one should keep in mind that only one of the states, LRS or HRS or any intermediate state, can be thermodynamically stable. Due to the nature of redox-based resistive switching, the other state(s) must be metastable. They are frozen-in after a kinetically fast (i. e. temperature- and/or field-accelerated) switching event. The reason for only *one* state being the thermodynamically stable state is due to the fact that there may be only *one* arrangement of ions and atoms which has the lowest free energy. This is different to ferroelectric and ferromagnetic systems in which

states with opposite polarization direction may both by thermodynamically stable. It should be mentioned that in ReRAM cell, both states may be metastable, e. g. characterized by a frozen-in enrichment or a frozen-in depletion of oxygen vacancies in front of an electrode interface.

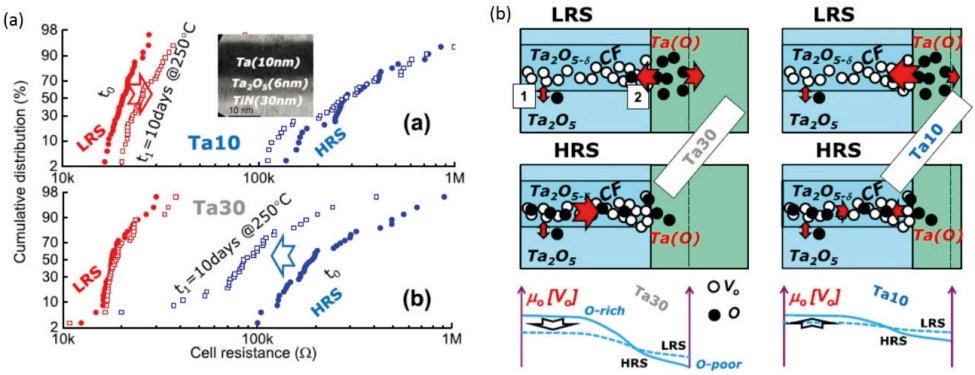


Fig. 20: *(a)Impact of the Ta thickness on the retention of LRS and HRS of TiN/Ta_2O_5/Ta cells.(b) Sketch of the explanation for the difference in the stability of HRS and LSR observed in (a), including a sketch of the oxygen chemical potential along the filament [35].*

Based on simulations of the *I-V*-characteristics and retention times, one finds that the retention failure mechanism for the LRS is based on the rupture of conducting filaments caused by re-oxidation due to oxygen diffusion from the side [36], [37] or along the vertical direction [38]. Recent studies on the technologically most relevant systems of HfO_{2-x} [39] and Ta_2O_{5-x} [35], however, reveal that extremely high retention times can be achieved with oxidizable electrodes or certain interlayers (the so-called oxygen scavenging layer). Phenomenologically, this finding was attributed to the stability of certain oxygen distributions in the layer stack and filaments with sufficient oxygen vacancy concentrations to be stable against re-oxidation. It is interesting to note that the data retention in TiN/Ta_2O_5/Ta cells crucially depends on the Ta thickness which results in a different oxygen vacancy distribution at the Ta/Ta_2O_5 interface as previously discussed. While for the 10nm thick Ta electrode, the LRS shows a significant resistance drift towards the HRS, the LRS is stable for the 30nm thick Ta electrode and the HRS drifts towards the LRS over time (figure 20(a)) [35]. As sketched in figure 20(b), the observation might be attributed to the difference in the chemical potential of oxygen along the filament, which originates from the different density of oxygen solved in the Ta electrodes with different thicknesses. This results in a thermodynamically stable LRS for the 30nm thick Ta electrode and a stable HRS for the 10nm thick Ta electrode.

Besides consideration about the thermodynamic stability of different states, an alternative approach is to increase the constraints for reoxidation to the system by kinetically hindering the oxygen back-diffusion. Noman *et al.* showed that in the absence of internal electric fields, $SrTiO_3$ cannot exhibit fast switching and long retention times simultaneously [40]. In fact, retention failure after short times was reported for the LRS in homogeneous single crystalline $SrTiO_3$ [41]. Polycrystalline and single crystalline $SrTiO_3$ films with considerable amounts of extended defects, on the other hand, exhibit much better retention behavior [41], [42] possibly induced by local variations of the diffusion constants. Furthermore, it has been observed that phase separation of $SrTiO_3$ into Sr-poor $SrTiO_3$ and SrO (figure 21(b)) results in a stabilization of the LRS (figure 21(a)) since oxygen diffusion is strongly hindered in the SrO layer [13]. As a result, a reoxidation of the oxygen deficient $SrTiO_3$ filament is prevented. A comparable

mechanism might take place in HfO_2/Hf cells, where the interface reaction results in the formation of an oxygen deficient HfO_{2-x} layer and a Hf electrode with a certain amount of dissolved oxygen. Since ab-initio calculation have shown that the oxygen vacancy mobility is strongly enhanced in oxygen-deficient HfO_{2-x}, this layer might act as sublayer where fast diffusion takes place [43]. On the other hand, the Hf layer may act as oxygen reservoir and as oxygen diffusion blocking layer [43] which prevents the reoxidation of the LRS. Besides these intrinsically formed bilayer systems, many groups intentionally grow bilayer systems containing one oxygen blocking layer such as Al_2O_3 in order to improve the device stability [13], [44].

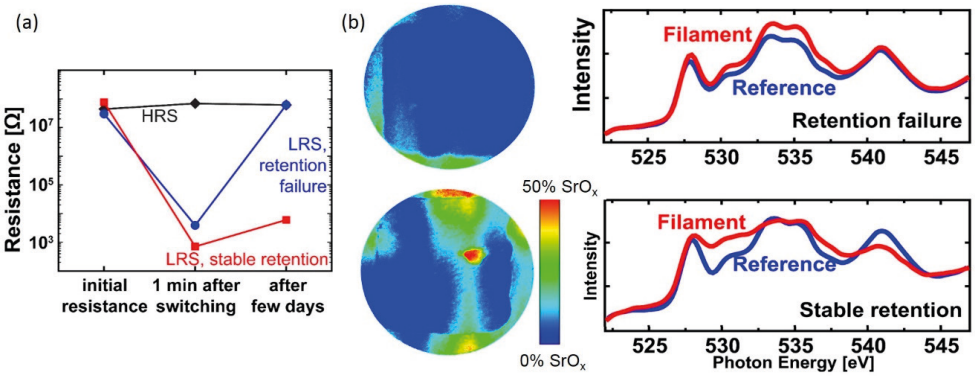


Fig. 21: *(a)Time dependence of the LRS and HRS for different $SrTiO_3$ thin film devices showing stable retention or LRS retention failure, (b) Spectroscopic investigation of devices with stable retention and retention failure: the right figures shows the comparison of the OK-edge which hints on the formation of SrO in the filament region of the device with stable retention. The PEEM false colour on the left depicts the spatial distribution of SrO on the device area. [13]*

7 Summary

The basic working principle of resistive switching in VCM cells is the field induced and temperature accelerated movement of donor type point defects which goes along with a valence change of the metal ion and results in a strong modulation of the electronic transport. Although it is in most cases assumed that oxygen vacancies are the dominant mobile defects, metal interstitials might be an alternative option for donor type defects responsible for some of the observed switching phenomena. During the first electrical biasing step, the so called electroforming procedure, a conducting filament is created which is subsequently locally interrupted during the switching process. Electroforming is based on ionic movement and an anodic oxidation, which goes along with the excorporation of oxygen and the local formation of a front of oxygen vacancies. However, detailed studies of the early stage of forming have shown that reversible electronic effects initiate the electroforming process and induce permanent changes in the ionic lattice in the later stage. Besides the creation of point defect oxygen vacancies, electroforming often goes the along with structural changes such as the formation of dislocations, stacking faults or phase changes. Resistive switching can to a certain extend be successfully described by the drift-diffusion of oxygen vacancies and the resulting modulation of the interface barriers.

Depending on the work functions of the two metal electrodes and the specific interface configuration, bipolar switching with two possible switching polarities (eightweise, counter-eightwise) or CS switching can be obtained. In the case of non-noble metal electrodes, however, the interface reaction between electrode and oxide is of key relevance, resulting in the formation of interface oxide layers on the one hand and in the formation of oxygen vacancies in the switching metal oxide on the other hand. In that case, interface reactions have to be considered as relevant for the switching process and might be superimposed with the drift-diffusion induced modulation of the interface barriers. In order to exhibit fast switching and long term stability at the same time, it is advantageous to employ bilayer systems (either intrinsically formed or intentionally grown) consisting of one sublayer with high oxygen vacancy mobility and an oxygen reservoir with a high kinetic barrier for the backdiffusion of oxygen.

References

[1] Waser, R., Dittmann, R., Staikov, G. & Szot, K. (2009) Advanced Materials, 21 (25-26), 2632-2663.

[2] Muenstermann, R. (2006) IFF Jülich, RWTH Aachen

[3] Waser, R., Bruchhaus, R. & Menzel, S. (2012) in: Nanoelectronics and Information Technology (3rd edition) (eds R. Waser), Wiley-VCH, pp. 683-710.

[4] Szot, K., Speier, W., Bihlmayer, G. & Waser, R. (2006) Nature Materials, 5 (4), 312-320.

[5] Havel, V., Marchewka, A., Menzel, S., Hoffmann-Eifert, S., Roth, G. & Waser, R. (2014) 2014 MRS Spring Meeting proceedings, MRS Online Proceedings Library.

[6] Yalon, E., Karpov, I., Karpov, V., Riess, I., Kalaev, D. & Ritter, D. (2015) Nanoscale, 7 (37), 15434-15441.

[7] Muenstermann, R., Menke, T., Dittmann, R. & Waser, R. (2010) Advanced Materials, 22 (43), 4819-4822.

[8] Dittmann, R. *et al.* (2012) Proceedings of the IEEE, 100 (6), 1979-1990.

[9] Lenser, Ch., Kuzmin, A., Purans, J., Kalinko, A., Waser, R. & Dittmann, R. (2012) Journal of Applied Physics, 111 (7), 76101.

[10] Kamaladasa, R. J. *et al.* (2013) Journal of Applied Physics, 113 (23), 234510/1-7.

[11] Szot, K., Rogala, M., Speier, W., Klusek, Z., Besmehn, A. & Waser, R. (2011) Nanotechnology, 22 (25), 254001/1-21.

[12] Kwon, J., Sharma, A. A., Bain, J. A., Picard, Y. N. & Skowronski, M. (2015) Advanced functional materials, 25 (19), 2876-2883.

[13] Bäumer, C. *et al.* (2015) Nature Communications, 6

[14] Sharma, A., Noman, M., Abdelmoula, M., Skowronski, M. & Bain, J. (2014) Advanced Functional Materials, 24, 5522–5529.

[15] Katharina Skaja, *et al.* (2015) Advanced Functional Materials, 25, 7154-7162.

[16] Sharath, S. U. *et al.* (2014) Applied Physics Letters, 104, 063502.

[17] Stille, S. *et al.* (2012) Applied Physics Letters, 100 (22), 223503/1-4.

[18] Govoreanu, B. *et al.* (2011) 2011 IEEE International Electron Devices Meeting - IEDM '11, IEDM Tech. Dig.

[19] Abbate, M. *et al.* (1991) Physical Review B, 44, 5419-5422.

[20] Park, G.-S. *et al.* (2013) Nature Communications, 4, 2382/1-9.

[21] Powell, C. J. & Jablonski, A. (2010) NIST Electron Inelastic-Mean-Free-Path Database-Version 1.2, National Institute of Standards and Technology, Gaithersburg, MD.

[22] Marchewka, A., Waser, R. & Menzel, S. (2015) 2015 International Conference On Simulation of Semiconductor Processes and Devices (SISPAD), Washington D.C, USA, 2015 International Conference On Simulation of Semiconductor Processes and Devices (SISPAD), Washington D.C, USA.

[23] Menzel, S. (2015) ECS Transactions, 69 (3), 19-32.

[24] Bruchhaus, R., Hermes, C. R. & Waser, R. (2011) MRS Online Proceedings Library, 1337, 73-78.

[25] Linn, E., Rosezin, R., Kügeler, C. & Waser, R. (2010) Nature Materials, 9 (5), 403-406.

[26] Schönhals, A. *et al.* (2015) Memory Workshop (IMW), 2015 IEEE International, Memory Workshop (IMW), 2015 IEEE International.

[27] Li, Z., Schram, T., Witters, T., Tseng, J., De Gendt, S. & De Meyer, K. (2010) Microelectronic Engineering, 87 (9), 1805-1807.

[28] Shibuya, K., Dittmann, R., Mi, S. & Waser, R. (2010) Advanced Materials, 22 (3), 411-414.

[29] Bertaud, T. *et al.* (2012) Applied Physics Letters, 101 (14), 143501/1-5.

[30] Scherff, M., Meyer, B., Hoffmann, J., Jooss, C., Feuchter, M. & Kamlah, M. (2015) New Journal of Physics, 17, 033011.

[31] Bertaud, T. *et al.* (2012) Symposium M on More than Moore - Novel Materials Approaches for Functionalized Silicon Based Microelectronics at Spring Meeting of the European-Materials-Research-Society (E-MRS), Strasbourg, FRANCE, E-Mrs 2012 Spring Meeting, Symposium M: More Than Moore: Novel Materials Approaches For Functionalized Silicon Based Microelectronics.

[32] Goux, L. *et al.* (2010) Applied Physics Letters, 97 (24), 243509.

[33] Lübben, M., Karakolis, P., Ioannou-Sougleridis, V., Normand, P., Dimitrakis, P. & Valov, I. (2015) Advanced Materials, 27 (40), 6202-6207.

[34] Wedig, A. *et al.* (2015) Nature Nanotechnology

[35] Goux, L., Fantini, A., Chen, Y. Y., Redolfi, A., Degraeve, R. & Jurczak, M. (2014) Ecs Solid State Letters, 3 (11), Q79-Q81.

[36] Ninomiya, T., Muraoka, S., Wei, Z., Yasuhara, R., Katayama, K. & Takagi, T. (2013) IEEE Electron Device Letters, 34 (6), 762-764.

[37] Ninomiya, T., Wei, Z., Muraoka, S., Yasuhara, R., Katayama, K. & Takagi, T. (2013) IEEE Transactions on Electron Devices, 60, 1384-1389.

[38] Gao, B. *et al.* (2008) Ieee International Electron Devices Meeting 2008, Technical Digest, 563-566.

[39] Chen, Y. *et al.* (2013) IEEE Transactions on Electron Devices, 60 (3), 1114-1121.

[40] Noman, M., Jiang, W., Salvador, P. A., Skowronski, M. & Bain, J. A. (2011) Applied Physics A: Materials Science & Processing, 102 (4), 877-883.

[41] Raab, N., Bäumer, C. & Dittmann, R. (2015) AIP Advances, 5, 047150.

[42] Dooho Choi, Dongsoo Lee, Hyunjun Sim, Man Chang, & Hyunsang Hwang, (2006) Applied Physics Letters, 88, 082904.

[43] Clima, S., Govoreanu, B., Jurczak, M. & Pourtois, G. (2013) Microelectronic Engineering

[44] Prezioso, M., Merrikh-Bayat, F., Hoskins, B. D., Adam, G. C., Likharev, K. K. & Strukov, D. B. (2015) Nature, 521 (7550), 61-64.

D 4 Switching Kinetics of Redox-based Resistive Memories

S. Menzel
Peter Grünberg Institut, PGI-7
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Switching kinetics: General considerations **3**
2.1 Limiting processes 3
2.2 Electric-field dependence of the switching kinetics 4
2.3 Temperature-dependence of the switching kinetics 5

3 Switching kinetics of VCM cells **6**
3.1 Analysis of the SET kinetics 7
3.2 Analysis of the RESET kinetics 8

4 Switching kinetics of ECM cells **12**
4.1 Analysis of the SET kinetics 12
4.2 Analysis of the RESET kinetics 14

5 Summary **16**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

Redox-based resistive switching devices based on the valence change mechanism (VCM) or the electrochemical mechanism (ECM) attract great attention for their utilisation in resistive redox-based random access memories (ReRAM) [1-4]. In ReRAMs the binary information is encoded as different resistance states, i.e. a low resistive state (LRS) and a high resistive state (HRS). ECM and VCM cells exhibit a bipolar operation scheme. They switch from the HRS to the LRS with one voltage polarity - the SET operation - and switch back with the opposite voltage polarity during the RESET operation. ReRAMs show very promising properties: non-volatility, fast device operation (< 10 ns) [5-7], high endurance (>10^{10} cycles) [8, 9] and good retention properties [10]. The write energy of these devices is in the rage of hundreds of fJ to a few pJ, which is higher than in conventional dynamic random access memory (DRAM). The non-volatility feature, however, leads to an overall lower energy consumption during operation. In order to meet the requirements for a competitive non-volatile memory, a ReRAM has to offer both fast switching and long-lasting read-disturb immunity. This means that upon excitation with only a few volts it has to switch as fast as nanoseconds. In contrast, upon applying a constant read voltage of a few hundred millivolts the resistance has to stay constant for up to ten years [3, 11]. Hence, the underlying physical processes that control the resistive switching mechanism should lead to a non-linearity of more than 15 orders of magnitude in time. In order to identify these processes, the ECM and VCM mechanisms are briefly reviewed.

A typical VCM cell consists of a metal oxide such as HfO_2 [12, 13], Ta_2O_5 [9, 14-16], TiO_2 [7, 17-20] or $SrTiO_3$ [21-23] sandwiched between an inert metal electrode, e.g. Pt or TiN, and an oxidisable one, e.g. Ta or Ti. In order to switch repetitively between the LRS and the HRS, an electroforming process is required in general. Thereby, the oxide is reduced by the extraction of oxygen via one of the electrodes under an applied voltage leaving behind oxygen vacancies. This results in the formation of a conducting filament consisting of a sub-stoichiometric oxide with a high oxygen vacancy concentration [3, 20, 24]. The resistive switching occurs in this filamentary region and is related to the movement of mobile donors such as oxygen vacancies. The local concentration of the mobile donors modulates the device resistance twofold. First, the local conductivity increases with increasing donor concentration. Second, the electro-static barrier at the metal/oxide interfaces decreases due to the Schottky effect, when the concentration of donors close to the interface increases [3, 19, 25-26][27]. In a VCM cell, a Schottky contact forms at the inert metal electrode and an ohmic contact forms with the oxidisable electrode. In this configuration the cell is set to the LRS, when the oxygen vacancies move to the Schottky contact and its barrier decreases enabling a high current injection. By reversing the polarity oxygen vacancies move away from this contact, the electrostatic barrier is reestablished and the cell resets. In addition to the movement of oxygen vacancies, oxygen exchange during switching could occur. Further details of the VCM switching processes are described in Chapter D3.

ECM cells consists of an active silver or copper electrode, an ion conducting layer, and an inert electrode (e.g. Pt). The switching mechanism is based on the electrochemical growth and dissolution of a silver or copper filament [28, 29]. When a positive voltage is applied to the active electrode, the electrode is oxidised and silver/copper ions are injected in the ion conducting layer. These ions migrate within the electric field toward the inert electrode where they are reduced. After a nucleation step a silver or copper filament forms, which grows to-

wards the active electrode until an electronic contact is established. Depending on the current limitation this could either be a galvanic contact or a remaining tunneling gap between filament tip and active electrode [30]. By reversing the voltage polarity, the electrochemical processes are reversed and the filament dissolves. Further details of ECM switching are described in Chapter D2.

This chapter focuses on the switching kinetics of ECM and VCM devices as described in [11]. First, it is described how physical and electrochemical process can limit the switching speed and affect the nonlinearity of the switching kinetics in general. Then, experimental observations and modeling results of ECM and VCM cells are discussed.

2 Switching kinetics: General considerations

In order to understand the nonlinear switching kinetics of ReRAMs one has to consider all electrochemical and physical processes that are involved in the resistive switching effect. According to the switching mechanisms described above these are: (i) Ion migration, (ii) electron-transfer (redox) reactions, and (iii) electrocrystallisation/nucleation. In a particular ECM or VCM cell all of these processes might be present. The slowest process, however, will limit the switching speed and determine the nonlinearity of the switching kinetics. In the following the degree of nonlinearity that results from these processes will be discussed.

2.1 Limiting processes

The ion migration processes can be mathematically described by the Mott-Gurney law for ion hopping

$$j_{\mathrm{hop}} = 2zeaf \exp\left(-\frac{\Delta W_{\mathrm{hop},0}}{k_{\mathrm{B}}T}\right)\sinh\left(\frac{aze}{2k_{\mathrm{B}}T}E\right). \tag{1}$$

Here, j_{hop} denotes the ionic current density, z the charge number of the hopping ion, e the elementary charge, f the jump attempt frequency, $\Delta W_{\mathrm{hop},0}$ the hopping barrier, k_{B} the Boltzmann constant, a the hopping distance, T the local temperature, and E the electric field. For high electric fields $E > 2k_{\mathrm{B}}T/(aze)$ an exponential relation between current density and electric field results. In contrast, the ionic current depends linearly on the applied electric field for $E < 2k_{\mathrm{B}}T/(aze)$. The electric field is defined as the voltage V that drops over a distance w according to $E = V/w$. The distance w can be the thickness of the switching layer or only a small part of it where the switching takes place.

The current j_{et} that results from electron-transfer reactions at the metal/ion-conducting layer interface is described by the Butler-Volmer equation

$$j_{\mathrm{et}} = j_{0,et}\exp\left(-\frac{\Delta W_{\mathrm{et}}}{k_{\mathrm{B}}T}\right)\left[\exp\left(\frac{(1-\alpha)ze}{k_{\mathrm{B}}T}\Delta\varphi_{\mathrm{et}}\right) - \exp\left(-\frac{\alpha ze}{k_{\mathrm{B}}T}\Delta\varphi_{\mathrm{et}}\right)\right] \tag{2}$$

and depends on the activation energy ΔW_{et}, the current prefactor $j_{0,\mathrm{et}}$, the charge transfer coefficient α, and the electron-transfer overpotential $\Delta\varphi_{\mathrm{et}}$. The first exponential term in Eq. (2) describes the oxidation reaction, whereas the second exponential term describes the reduction reaction. If the overpotential is zero, both processes occur at the same rate and the redox reac-

tion is in a dynamic equilibrium. For $\Delta\varphi_{et} > 0$ the oxidation reaction overweighs, whereas the reduction dominates for $\Delta\varphi_{et} > 0$.

Electrocrystallisation describes nucleation and crystal growth in electrochemical systems under influence of electric fields. In this process also a charge transfer is involved in the formation of a new phase. The nucleation time t_{nuc}, which is the required time to form a stable nucleus, depends exponentially on the nucleation overpotential $\Delta\varphi_{nuc}$ according to

$$t_{nuc} \propto \exp\left(\frac{\Delta W_{nuc}}{k_B T}\right)\exp\left(-\frac{(N_c+\alpha)ze}{k_B T}\Delta\varphi_{nuc}\right). \tag{3}$$

Here, ΔW_{nuc} is the nucleation activation energy and N_c gives the number of atoms that is required to form a stable nucleus.

Despite their different physical and electrochemical nature, all of these processes obey an Arrhenius-type law. Thus, all processes are exponentially enhanced when the temperature increases due to local Joule heating. In addition, the activation barrier can be lowered by a sufficiently high electric field, which results in an exponential dependence on the electric field. It should be noted that the minimum resulting effective barrier height is 0. Thus, the activation energy also contains the information on how many orders of magnitude in nonlinearity are achievable. To achieve more than 15 orders of magnitude at an ambient temperature of 300 K (400 K) an activation energy $\Delta W > 0.9$ eV (1.2 eV) is required [11]. This activation energy only gives the required amount of nonlinearity. To fulfill the 10 y retention requirement the activation energy has to be slightly higher [31].

2.2 Electric-field dependence of the switching kinetics

Due to the physical parameters the different processes exhibit different degrees of nonlinearity. In a pulse experiment, plotting the logarithm of the switching time $\ln(t_{sw})$ against the switching voltage V_{sw} would result in different slopes m. Thus, the analysis of the slopes can help to identify the limiting processes. According to Eqs. (1) to (3), the slopes

$$m_{hop} = -\frac{aze}{2k_B T w},\ m_{red} = -\frac{aze}{k_B T},\ m_{ox} = -\frac{(1-\alpha)ze}{k_B T},\ \text{and}\ m_{nuc} = -\frac{(N_c+\alpha)ze}{k_B T} \tag{4}$$

can be extracted for the ion hopping, reduction, oxidation and nucleation process, respectively. Instead of using the slope m for comparison of different processes, one can use the voltage increment

$$\Delta V_{10x} = -m^{-1}\ln(10) \tag{5}$$

that is required to accelerate the switching speed by a factor of 10. The voltage increment and the corresponding slope m_{exp} can be extracted from experiment. Using Eq. (4) and Eq. (5) and assuming a single limiting process, the required physical parameters to achieve this increment can be calculated and checked for physical meaningfulness. Here, two cases are of special interest: i) If $|m_{exp}\cdot k_B T/(ze)| > 1$, only assuming the nucleation process as limiting factor results in physically meaningful parameters. ii) For the ion hopping process, a physically reasonable value for the ratio a/w should be smaller than 1/5. In that case the voltage drops over a distance of not less than w = 1.5-2.5 nm for a reasonable hopping distance of 0.3-0.5 nm. Thus, if $|m_{exp}\cdot k_B T/(ze)| > 1/5$, ion hopping can be excluded as the only limiting process.

Assuming some reasonable limiting parameters, which are given in the figure caption, the different slopes are plotted as normalised switching time versus applied voltage in Figure 1 [11]. The nucleation process shows the highest nonlinearity followed by the electron-transfer reaction and the ion hopping process. When more than a single process is present in a specific device, the slowest one will determine the slope in the *t-V* diagram. The process with the steepest slope will most likely determine the slope at low voltages, whereas a process with a flatter slope will limit the switching speed at higher voltages. Overall, one would expect that the slope in the *t-V* diagram is flattening out when the nonlinearity is only achieved by electric field.

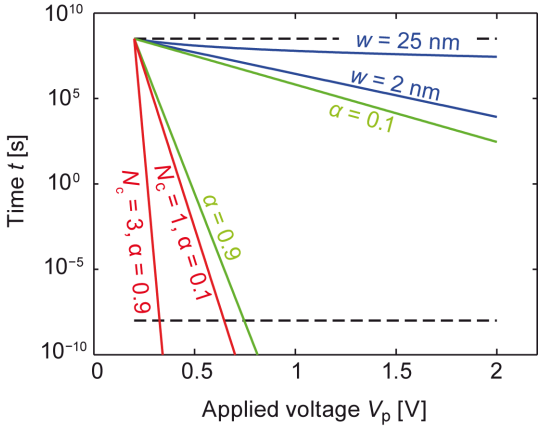


Fig. 1: *Illustration of the nonlinearity in the switching kinetics obtained for electric-field enhanced nucleation (red), electron-transfer reaction (green), and ion migration (blue) in the limiting scenarios explained below. The different processes cover a different range of slopes in the t-V diagram. For the ion migration curves a = 0.3 nm, and w = 2 nm and 25 nm are chosen as lower and upper limit. The charge transfer coefficient is chosen in a range between* $0.1 \leq \alpha \leq 0.9$. *For the nucleation* $N_c = 1$, $\alpha = 0.1$ *and* $N_c = 3$, $\alpha = 0.9$ *are used. For all lines z = 2 is assumed. Figure reprocuded from [11].*

2.3 Temperature-dependence of the switching kinetics

As the resistive switching in ReRAM devices takes place in a filamentary region, Joule heating is expected to occur for significantly high dissipated power P_{el}, i.e. higher than a few microwatts. The local temperature can be estimated using

$$T = T_0 + R_{th} P_{el} = T_0 + R_{th} I(V) V , \tag{6}$$

where T_0 is the ambient temperature and R_{th} the equivalent thermal resistance. The latter depends on the thermal conductivities of the filament, the surroundings and the electrodes, and the geometry of the filament. According to Eq. (6) a strong power-dependence has to be expected. In the following, the temperature-acceleration of the ion hopping process is discussed. The conclusions, however, can be also extended to the other processes.

Figure 2 shows the interdependence of non-linearity and dissipated electrical power for two different current-voltage scenarios: (i) an ohmic behaviour with $I = V/R$, and (ii) a diode-like behaviour, i.e. $I = I_0 \cdot (\exp(V/V_0)\text{-}1)$. The parameters are given in the figure caption. The local Joule heating effect clearly increases the switching speed. For low voltages, the ohmic (blue) and the diode-like (red) scenarios equal the constant temperature case (black). As soon as Joule heating sets in, the slope becomes steeper than in the case of sole voltage/field acceleration with constant temperature. For the ohmic behaviour, Joule heating sets in at lower voltages than for the diode-like behaviour as the current is higher at low voltages. The crossing point in *t-V* diagram marks the point where the *I(V)* pair of values are identical for both scenarios. From the comparison of both scenarios it appears that the nonlinearity is highly dependent on the nonlinearity of the *I-V* relation. It is remarkable that the switching kinetics differ strongly in the *t-V* diagram but are almost similar in the *t-P* diagram. This illustrates the strong power-dependence in the chosen scenario. The small difference in the *t-P* diagram can be related to the voltage/field acceleration of the ion hopping process. For $P < 10$ µW the temperature increase is below 15 K and the influence of Joule heating is small. To achieve the same power, however, a higher voltage has to be applied in the diode-like case than in the ohmic case. As a result the switching speed is slightly higher. In order to prove that the switching kinetics are power-dependent in a real ReRAM device it is useful to compare the *t-P* curves for different programmed initial resistances. If only one process is limiting the switching speed, coinciding *t-P* curves should result for different initial resistances. This behaviour might change if several processes limit the switching speed.

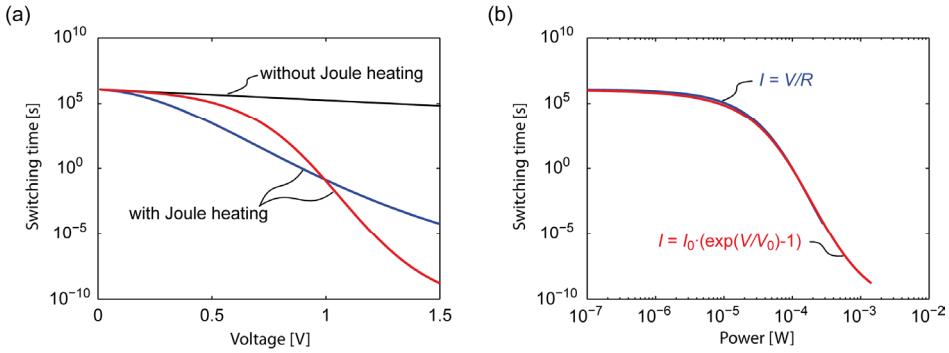


Fig. 2: *(a) Illustration of the switching time vs. applied voltage calculated without Joule heating (black solid line) and with Joule heating assuming a linear I-V relation (blue solid line) and a diode like I-V behaviour (red solid line). The corresponding switching time vs. dissipated power plot is shown in (b) using the same colour code. The parameters used are: a = 0.5 nm, ΔW_{hop} = 1 eV, V_0 = 0.25 V, I_0 = 2.38 µA, R = 10 kΩ, $R_{th} = 1.25 \cdot 10^6$ K/W.*

3 Switching kinetics of VCM cells

The SET and RESET switching kinetics have been investigated for different device stacks using pulse experiments. Whereas systematic studies of the SET kinetics have been presented frequently, only few RESET kinetics study are available. One reason might be the nature of the switching event. While an abrupt current jump in the current transient indicates the SET time, the RESET transition is gradual. Thus, it is not easy to analyse the RESET kinetics sys-

tematically. The RESET time also depends on its definition, which makes a fair comparison of different studies hardly possible.

3.1 Analysis of the SET kinetics

Figure 3 shows the published switching kinetics data of TiN/HfO_x/TiN [32, 33], TiN/Ti/HfO_x/TiN [34], TiN/HfO_x/AlO_x/Pt [35], Ti/HfO_x/Pt [36], Pt/TiO_x/Pt [37], Pt/TaO_x/Ta [38], and Pt/$SrTiO_3$/Ti [39] devices, as compiled in [11]. The observed inverse slopes are all very similar in a range of ΔV_{10x} = 40-240 mV/dec. Moreover, most of the devices show only a single slope except for the TiN/HfO_x/TiN cells of Ielmini and co-workers (red open squares) [33], the HfO_x data published by Cao and co-workers (red filled squares) [36], and the $SrTiO_3$ data of Fleck and co-workers (black open squares) [39]. However, most of the studies only cover less than five orders of magnitude in switching time and another slope might appear in the non-studied voltage regimes. According to the switching mechanism discussed above the migration of double-positively charged oxygen vacancies is supposed to limit the switching speed. In order to explain the slopes by pure field-acceleration the voltage needs to drop over no more than 0.6 nm – 2.1 nm. This value seems to be quite low. In fact, the transient currents prior to the abrupt SET transition are typically higher than a few µA and hence Joule heating should occur. Thus, a combination of electric field and temperature acceleration is the most likely scenario. In fact, the data of Cao and co-workers (red filled squares) and Fleck and co-workers (black open squares) show a flattening at lower voltages, which indicates the role of Joule heating.

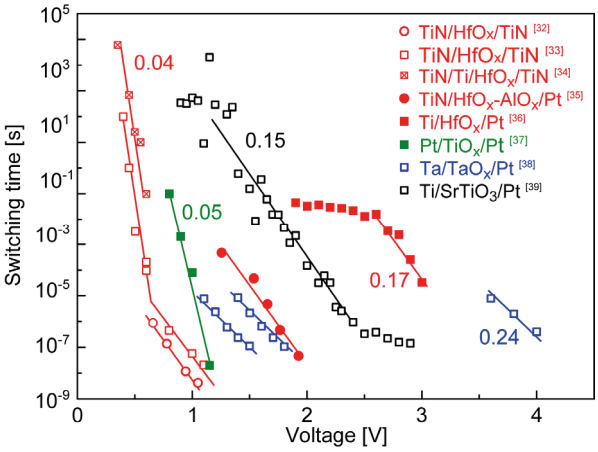


Fig. 3: *Switching kinetics data for VCM cells of hafnium oxide (red) [32-36], titanium oxide (green) [37], tantalum oxide (blue) [38], and strontium titanate (black) [39]. A specific slope ΔV/dec(t) of each oxide material can be identified in a narrow range. Shifts along the voltage axis for same species are related to an increase of the oxide thickness. For HfO_x and $SrTiO_x$-based cells regimes with different slopes are observed. Figure reproduced from [11].*

If temperature-acceleration dominates, a clear power-dependence of the switching time on the dissipated power should be expected. Nishi et al. analysed the switching kinetics data for two 5 nm thick TaO_x-based VCM cells A and B with different high resistive state [38]. As illustrated in Figure 4 the two data sets exhibit almost the same slope but are shifted by about 0.3 V [11]. In contrast, the data sets coincide when plotted against the dissipated power during switching. As discussed above this is a clear indication for the dominant role of Joule heating in explaining the nonlinear switching kinetics. A similar behaviour was also demonstrated for measurements at room temperature and at 85°C [38]. The analysis of this data reveals a difference in the *t-V* diagram, but the data coincides in the *t-P* diagram.

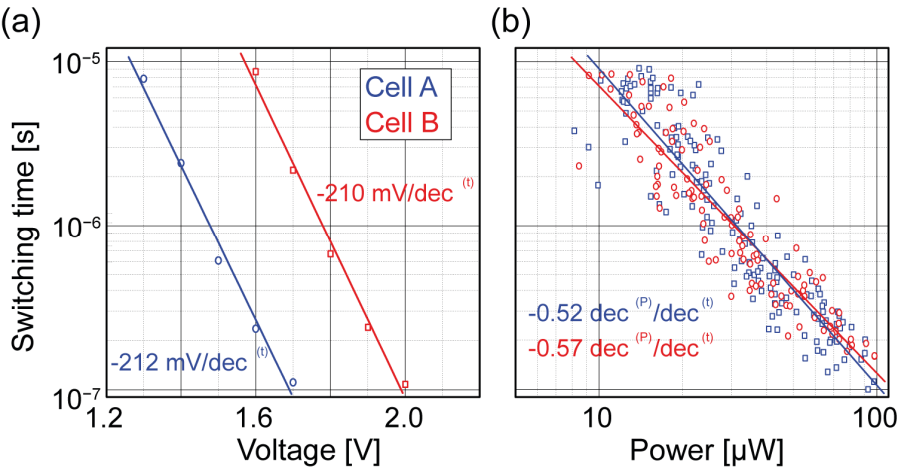


Fig. 4: *(a) SET switching time vs. applied voltage for different TaO_x-based VCM cells A (blue) and B (red). (b) SET switching time plotted against the corresponding SET power extracted from the SET transients for both cells. While there is parallel shift in the t-V data, the t-P data coincide. Data are adopted from Nishi et al. [38]. Figure reproduced from [11].*

Further evidence for the importance of Joule heating is given by several groups using a combined experimental and simulation approach. In 2011, Menzel et al. used an electro-thermal model to investigate the switching kinetics of $SrTiO_3$-based VCM cells [40]. By comparison of the experimental data with different simulation scenarios temperature-accelerated ion hopping could be identified as the dominating process in explaining the nonlinear switching kinetics. Furthermore, Ielmini et al. developed a VCM switching model based on radial filamentary growth driven by an Arrhenius type law [33]. The local temperature increase was calculated according to Eq. (6). With this model they were able to reproduce the switching kinetics data of the TiN/HfO_x/TiN device shown in Figure 3 as red squares. Fleck et al. derived an analytical approach for determining the switching time that is based on temperature-accelerated ion hopping [39]. The $SrTiO_3$-based VCM cell shown as black open squares in Figure 3 could be accurately described with this model.

As the resistive switching is explained in terms of ion migration, the retention of the programmed devices states is described by ion redistribution due to diffusion, at least, as no other process stabilises the filament. Noman et al. demonstrated by means of drift-diffusion simulations that the activation energy of ion diffusion should exceed 1.02 eV in order to achieve a 10 year retention [31].

3.2 Analysis of the RESET kinetics

In contrast to the SET transition the RESET transition in VCM devices is typically gradual. This phenomenon can be used to program different intermediate resistance states by changing either the RESET “stop” voltage in sweep measurements [41-44] or the reset voltage amplitude in pulse experiments [35, 45-46]. Only a few studies on the RESET dynamics have been published so far [35, 46-48]. Thus, this section focuses on the explanation of the gradual RE-

SET phenomenon. In 2015, Marchewka et al. developed a 2D dynamic model of non-isothermal drift-diffusion transport to analyse the RESET transition [26]. This model will be described in the following.

The model considers an axisymmetric model geometry as shown in Figure 5 [26]. It comprises a 5 nm thick TaO_x film sandwiched between a Ta/Pt top electrode and Pt bottom electrode, which is deposited on a SiO_2 substrate. The switching occurs in a filamentary region within the TaO_x film with high oxygen vacancy concentration. It is assumed that the TaO_x forms an ohmic-like contact with the Ta electrode and a Schottky-like contact with the bottom Pt electrode. In order to switch the device from the LRS to the HRS the oxygen vacancy concentration needs to be depleted at the Schottky-like contact.

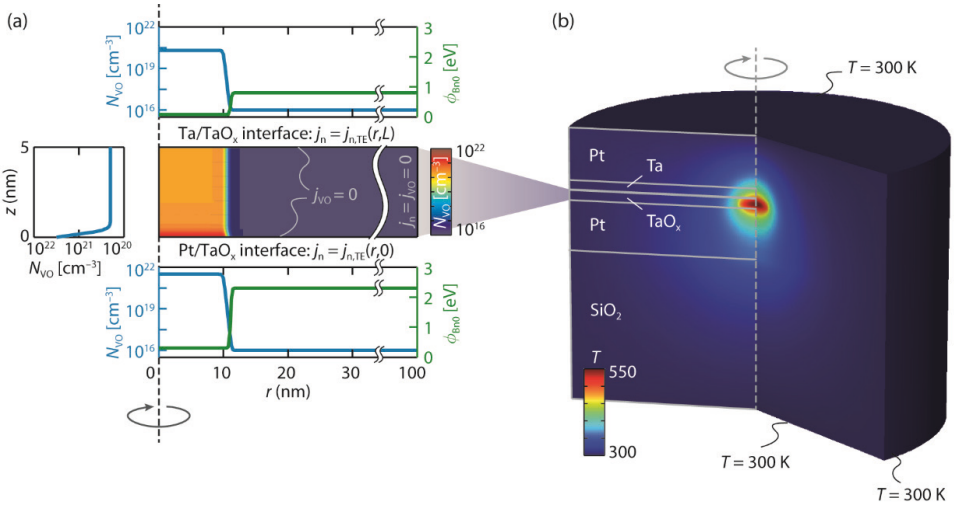


Fig. 5: *Model geometry. (a) Computational domain of the TaO_x layer with initial and boundary conditions used in the drift-diffusion simulation. Center: Map of the initial oxygen-vacancy distribution inside the TaO_x layer. The boundary conditions for the electronic and ionic currents are indicated at the domain boundaries. Top: Radial initial donor distribution $N_{VO}(r,L)$ and barrier heights $\phi_{Bn0}(r,L)$ at the Ta/TaO_x interface. Bottom: Radial initial donor distribution $N_{VO}(r,0)$ and barrier heights $\phi_{Bn0}(r,0)$ at the Pt/TaO_x interface. Left: Initial donor distribution $N_{VO}(0,z)$ in the filament center. (b) Computational domain comprising the layer stack of 75 nm SiO_2, 25 nm Pt, 5 nm TaO_x, 5 nm Ta, and 25 nm Pt used for the temperature calculation, along with the boundary conditions for the heat equation. A typical temperature distribution is shown as an example. Figure reproduced from [26].*

Neglecting the minority carriers and assuming that the oxygen vacancies are twofold ionizable the Poisson equation is expressed as

$$\Delta\psi = -\frac{e}{\varepsilon_0\varepsilon_r}\left(n - N_{VO}^{+} - 2N_{VO}^{2+}\right). \tag{7}$$

It is solved along with the steady-state continuity equation for electrons

$$-\nabla\left(-\mu_n n\nabla\psi + D_n\nabla n + nD_{Tn}\nabla T\right) = \pm\frac{\partial j_{n,tunnel}}{\partial z}, \tag{8}$$

the time-dependent continuity equation for the doubly ionised oxygen vacancies

$$\frac{\partial N_{VO}^{2+}}{\partial t} - \nabla\left(\mu_{VO} N_{\mathrm{VO}}^{2+} \nabla\psi - D_{\mathrm{VO}} \nabla N_{\mathrm{VO}}^{2+} - N_{\mathrm{VO}}^{2+} D_{\mathrm{TVO}}\right) = -R_{\mathrm{VO},2}\,, \tag{9}$$

the rate equation for the immobile singly ionised oxygen vancancies

$$\frac{\partial N_{\mathrm{VO}}^{+}}{\partial t} = -R_{\mathrm{VO},1}\,, \tag{10}$$

the rate equation for the immobile neutral oxygen vacancies

$$\frac{\partial N_{\mathrm{VO}}^{0}}{\partial t} = -R_{\mathrm{VO},0}\,, \tag{11}$$

and the heat transfer equation

$$-\nabla\left(k_{\mathrm{th}} \nabla T\right) = jE\,. \tag{12}$$

In Eqs. (7)-(12), ε is the permittivity of the oxide, ψ the potential, n the electron concentration, $N_{\mathrm{VO}}{}^{+}$ ($N_{\mathrm{VO}}{}^{2+}$) the concentration of singly (doubly) ionised oxygen vacancies, μ_{n} (μ_{VO}) the electron (oxygen-vacancy) mobility, D_{n} (D_{VO}) the electron (oxygen-vacancy) diffusion coefficient, D_{Tn} (D_{TVO}) the electron (oxygen-vacancy) thermal diffusion coefficient, $N_{\mathrm{VO}}{}^{0}$ the neutral oxygen vacancy concentration, k_{th} the thermal conductivity, T the local temperature, j the local current density and E the electric field. The right hand side of Eq. (8) describes a local generation/recombination rate due to electron tunnelling through the contact potential barriers. $R_{\mathrm{VO},2}$, $R_{\mathrm{VO},1}$ and $R_{\mathrm{VO},0}$ represent the reaction rates that are derived from the laws of mass action along with oxygen-vacancy ionisation statistics [26].

The electron transport across the metal-oxide contact has two different contributions: electron tunnelling and thermionic emission. The electron tunnelling contribution through a barrier with energy minimum W_{min} and energy maximum W_{max} is calculated according to

$$j_{\mathrm{n,tunnel}} = \frac{A^{*}}{k_{\mathrm{B}}^{2}} \int_{W_{\mathrm{min}}}^{W_{\mathrm{max}}} \mathcal{T}\left(W_z\right) N\left(W_z\right) dW_z\,. \tag{13}$$

Here, A^{*} is the effective Richardson constant, $\mathcal{T}(W_z)$ is the transmission coefficient obtained from the Wentzel-Kramers-Brillouin (WKB) approximation and $N(W_z)$ is the supply function. The latter describes the supply with carriers and is derived by integration of the occupancy functions on both sides of the barrier. For thermionic emission, the transmission coefficient is 1 and the current is obtained by integrating over all energies from the conduction band edge W_{c} at the contact interface to infinity according to

$$j_{\mathrm{n,TE}} = \frac{A^{*}}{k_{\mathrm{B}}^{2}} \int_{W_{\mathrm{c}}}^{\infty} 1 \cdot N\left(W_z\right) \mathrm{d}W_z\,. \tag{14}$$

This set of equations is complemented by appropriate boundary conditions as indicated in Figure 5 and further outlined in [26].

This model was applied to analyse the gradual RESET transition in TaO_x-based VCM cells [26]. The transient currents upon voltage pulses with a rise time of 2 ns, a duration of 1 µs and different voltage amplitudes were simulated. Figure 6(a) shows the simulated current transients (in colour) for pulses with amplitude −1.3 V, −1.4 V, −1.5 V and −1.6 V compared to experimental data. The model reproduces the experimentally observed transient behaviour very well. The point C in each transient marks the decay time $\tau_{50\%}$ when the current drop is half of the total current drop occurring during the pulse, i.e. the difference in currents in point

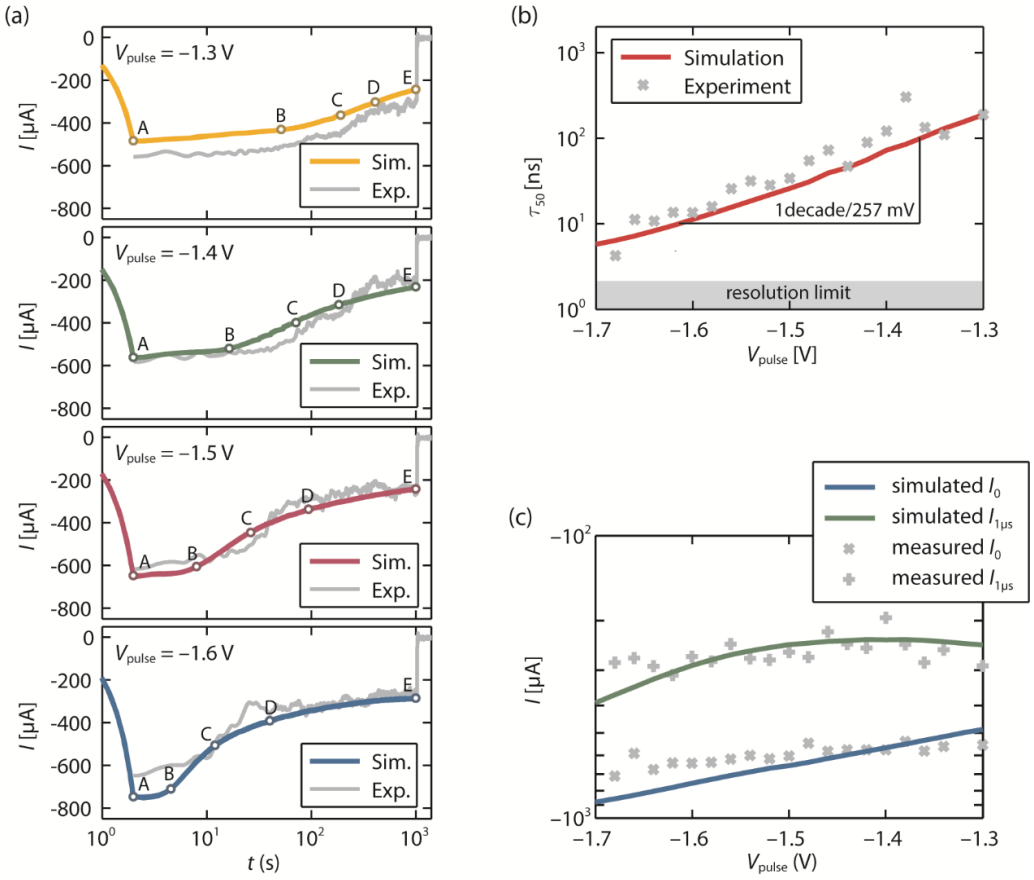


Fig. 6: *Comparison between simulation and measurement of (a) transient currents for pulse voltages of −1.3 V, −1.4 V, −1.5 V and −1.6 V, (b) 50% decay times as a function of pulse voltage, (c) current I_0 at the beginning of the pulse and current $I_{1µs}$ at the end of the pulse as functions of pulse voltage. Figure reproduced from [26].*

A and E. As shown in Figure 6(b), the decay time depends exponentially on the voltage pulse amplitude, which illustrates the nonlinearity of the RESET switching kinetics [26]. To accelerate the switching speed by one order of magnitude a voltage increment of 257 mV is required, which is larger than for the SET operation. The measured decay time is reproduced well by the simulation model. In addition, the simulated currents at points A and E are in good agreement with the experimental data (Figure 6(c)). By analysing the simulated transient current contributions, temperature and concentration profiles, the origin of the gradual RESET transition could be identified. At the beginning of the pulse the oxygen vacancy concentration is approximately homogeneous. Due to the high current density, local Joule heating occurs and the ions drift within the applied electric field toward the ohmic electrode. As the ions redistribute the potential barrier at the Schottky-like contact is increased and thus the

current decreases. This leads to a decrease in temperature. Thus, the ionic current is reduced and the driving force for the RESET transition is lowered. In addition, a concentration gradient builds up and ion diffusion sets in that counteracts the ion drift. To conclude, the gradual nature of the RESET transition can be explained by the temperature-accelerated oxygen-vacancy motion with the drift and diffusion processes approaching an equilibrium situation, combined with a moderate sensitivity of the current response to the induced contact barrier changes [26]. A 1D variant of this simulation model including barrier lowering due to the Schottky effect has been used to simulate the quasi-static *I-V* characteristics, the SET transients, and complementary switching behaviour [25]. Some of those simulation results are shown in chapter D3.

4 Switching kinetics of ECM cells

Similar as for the VCM cells, SET switching kinetics of ECM cells have been studied quite frequently, but RESET kinetics studies are scarce. Thus, only the SET kinetics will be compared for different device stacks according to [11]. The RESET switching kinetics will be explained for GeS_x-based ECM cells [49].

4.1 Analysis of the SET kinetics

Figure 7 shows the compilation given in [11] of the published SET switching time data as a function of applied voltage [50-58]. From the data three different groups were identified that show similar behaviour: i) primary solid electrolytes, ii) secondary solid electrolytes and iii) untypical solid electrolytes. The primary electrolytes comprise Cu_2S, Ag_2S, AgI or $RbAg_4I_5$, where the metal species of the composition is intrinsically present as cations. In contrast, the secondary electrolytes as GeS_x or $GeSe_x$ are well known ionic conductors for Ag or Cu cations, which are extrinsically delivered by doping during processing and/or by in-diffusion from the electrode after deposition. In the untypical solid electrolytes like a-Si and Ta_2O_5, a counter charge is required in order to inject Cu or Ag cations. For Cu/Ta_2O_5 and Cu/SiO_2 systems it has been shown that the counter charges are possibly OH^- ions provided by residual water within the thin film [59-61]. Thus, water serves as an electrolyte in these systems. For a nanoporous Cu/HfO_2 ECM cell it was shown that the solvent used as electrolyte influences the switching characteristics [62].

In contrast to the VCM data in Figure 3, several slopes appear for the individual ECM data sets in the *t-V* diagram (Figure 7). Thus, different processes limit the switching speed in different voltage regimes. With increasing voltage the slope flattens out, which indicates that the nonlinearity of the switching kinetics is dominated by electric-field acceleration of the underlying processes. Analysing the different slopes, which are given as inset in Figure 7, suggests that nucleation limits the switching speed at very low voltages, followed by electron-transfer reactions and ion hopping. A detailed discussion of the slopes is given in [11].

The three different groups differ greatly in the switching speed. This difference can be attributed to the concentration of Cu or Ag cations in the thin film. In the primary electrolytes the concentration is very high. As the electron-transfer and the ion hopping process are proportional to the ion concentration, the ionic current density is very high and the filament can be built up quickly. In the secondary solid electrolytes the concentration of Cu or Ag ions is determined for example by temperature-assisted in-diffusion from the active metal electrode

after deposition. It has been shown for a Ag:GeS_x system that due to this process the cation concentration is highly thickness-dependent [49]. The SET switching speed increases by seven orders of magnitude when the thickness is reduced from 100 nm to 20 nm. A similar trend exhibits the data from Palma and co-workers shown as red triangles in Figure 7 for high voltages [54]. For the unconventional solid electrolytes the ion concentration depends on the solvent that serves as solid electrolyte [62]. By using a solvent with a high cation solubility the switching voltage could be successfully reduced.

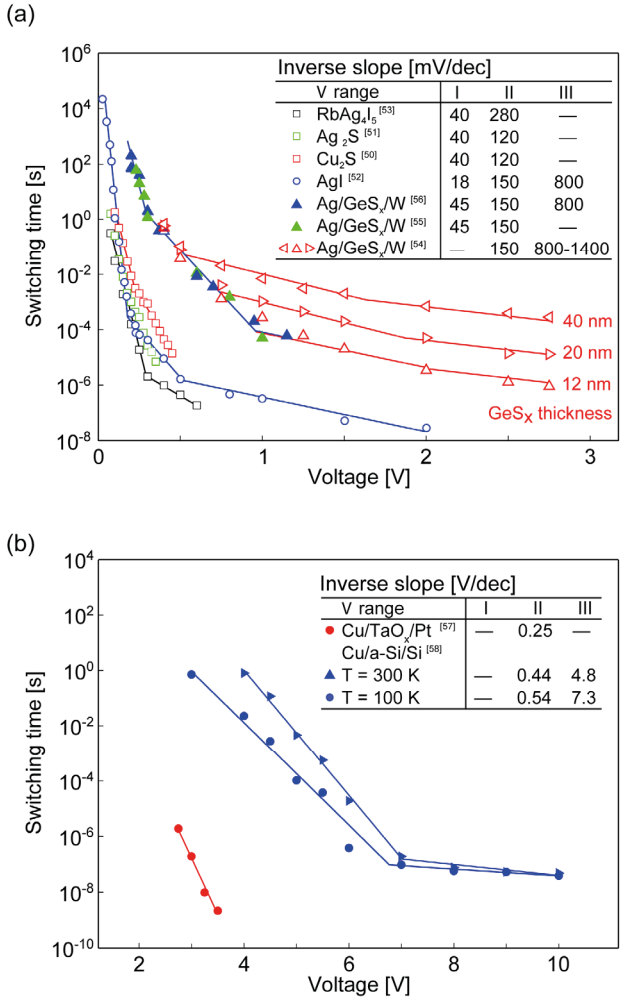


Fig. 7: *Switching kinetics (log (t)–V) of ECM cells showing different inverse slopes in the specific voltage regimes "low" (I), "medium" (II). "high" (III) for (a) primary solid electrolytes Ag_2S [51], Cu_2S [50], $RbAg_4I_5$ [53], and AgI [52] as well as secondary solid electrolytes Ag-GeS_x [54-56], and (b) untypical solid electrolytes Cu/TaO_x/Pt [57] and Cu/a-Si/Si [58]. Figure reproduced from [11].*

In order to explain the SET switching kinetics of the AgI-based ECM cells shown as open blue circles in Figure 7, Menzel and co-workers developed a 1D simulation model [52]. In this model a cylindrical filament is considered that grows within a switching layer of thickness L. Figure 8(a) shows the equivalent circuit diagram of the model [52]. It includes the electron-transfer reactions at the metal/switching layer interfaces, i.e. η_{ac} and η_{fil}, and the ion hopping process η_{hop} according to Eq. (2) and Eq. (1), respectively. The filamentary growth velocity is described in terms of the tunnelling gap x between the filament and the active electrode according to [30, 52]

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{M_{\mathrm{Me}}}{ze\rho_{\mathrm{m,Me}}} J_{\mathrm{ion}} . \tag{15}$$

Here, M_{Me} is the atomic mass and $\rho_{\mathrm{m,Me}}$ the mass density of the deposited metallic species. The ionic current density J_{ion} is determined using Eq. (1). The electron tunnelling current I_{Tu} is calculated according to Simmons [63] by

$$I_{\mathrm{Tu}} = C \frac{3\sqrt{2m_{\mathrm{eff}}\Delta W_0}}{2x} \left(\frac{e}{h}\right)^2 \exp\left(-\frac{4\pi x}{h}\sqrt{2m_{\mathrm{eff}}\Delta W_0}\right) A_{\mathrm{fil}} V_{\mathrm{Tu}} . \tag{16}$$

In Eq. (16) m_{eff} denotes the effective electron mass, ΔW_0 the tunnelling barrier height, h Planck's constant, A_{fil} the filament cross-section area, and C a fitting constant [52, 64]. The filamentary growth can be simulated by solving the ordinary differential equation Eq. (15). Prior to the filamentary growth the nucleation time is calculated according to Eq. (3).

This model has been applied to investigate the switching dynamics of the AgI-based ECM cells shown in Figure 7. As shown in Figure 8(b) the simulation model can reproduce the experimental data for different temperatures in the complete voltage range. From the analysis of the transient simulation data the limiting process in the different voltage regimes can be extracted. For low voltages the nucleation takes up most of the switching time and thus limits the switching speed (cf. Figure 8(c)). In the intermediate voltage range the nucleation time is negligible and the filament growth determines the switching time as shown in Figure 8(d). As the hopping overpotential is almost zero, the electron-transfer reaction limits the switching speed. For voltages higher than 1 V the hopping overpotential is in the range of the electron-transfer overpotentials (Figure 8(e)) and thus the switching speed is determined by electron-transfer reactions and ion hopping.

This model has been used to explain the switching dynamics of a Ag:GeS_x based ECM cell including the aforementioned thickness dependence of the switching time [49].

4.2 Analysis of the RESET kinetics

The RESET switching kinetics of a Ag/GeS_x/Pt cell have been investigated using the simulation model described above. The simulated RESET times are plotted against the applied voltage along with the experimental data in Figure 9. The simulation model reproduces the experimental data very well. By analysing the simulated transient data, the limiting processes could be identified. For low voltages the electron-transfer reaction limits the switching speed and at higher voltages a combination of electron-transfer and ion hopping processes determines the switching time [49]. In this study, it was also demonstrated that the RESET switching time is independent of the programmed LRS.

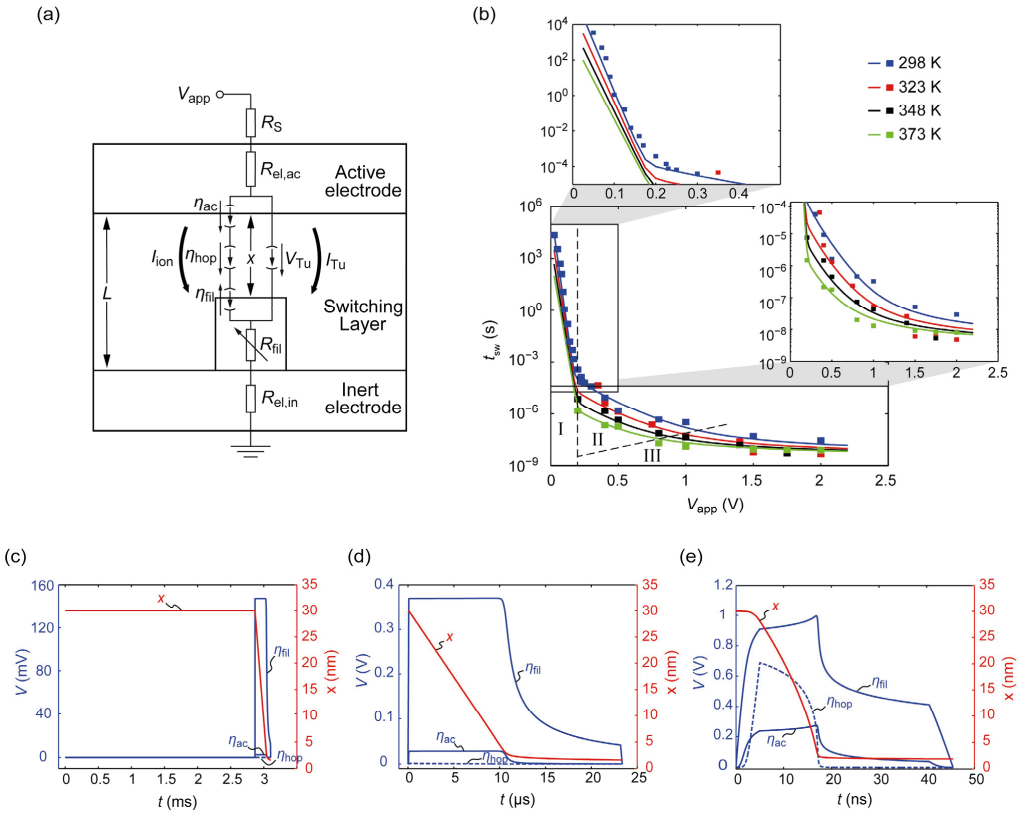


Fig. 8: *(a) Schematic of the ECM switching model with an equivalent circuit diagram. A switching layer of thickness L is sandwiched between the active top electrode and the inert bottom electrode. A cylindrical filament grows within switching layer and modulates the tunneling gap x between the filament and the active electrode. In the switching layer both ionic and electronic current paths are present. (b) Pulsed SET switching kinetics of a AgI-based ECM cell for different ambient temperatures T = 298 K (blue), 323 K (red), 348 K (black) and 373 K (light green). The simulated data are displayed using solid lines and the experimental data using squares. I, II and III mark the nucleation limited, the electron-transfer limited and the mixed control regime, respectively. The corresponding transient overpotentials (blue) and tunneling gaps (red) are shown for an applied voltage of (c) 0.15 V representing the nucleation limited regime, (d) 0.4 V representing the electron-transfer limited regime and (e) 2V, which corresponds to the mixed control regime. In (c)-(d) the hopping overpotential is illustrated with blue dashed lines and the electron-transfer overpotentials with blue solid lines. From [52]. - Reproduced by permission of the PCCP Owner Societies.*

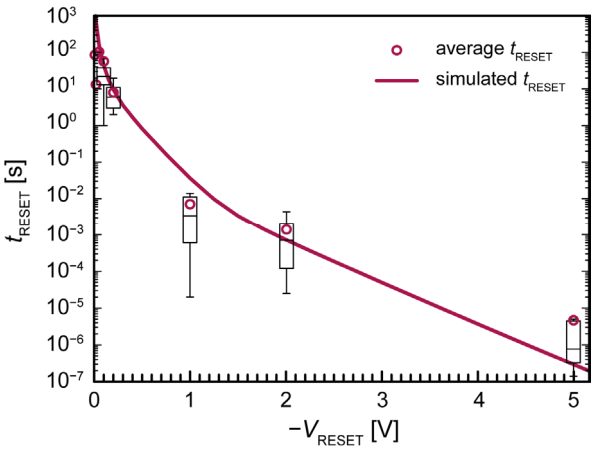


Fig. 9: *Pulse measurements of 2 µm* x *2 µm microcrossbar memory cells with 70 nm GeSx. (b) Semilog plot of RESET time t_{RESET} versus RESET voltage V_{RESET}. Error bars for $V_{RESET} > -100$ mV have been removed for clarity. Simulation result is given as solid line. Figure reproduced from [49].*

5 Summary

In this chapter the switching kinetics of redox-based resistive switching devices were discussed. The involved electrochemical and physical processes can be either electric field/voltage enhanced or accelerated by a local increase in temperature due to Joule heating. If only electric field/voltage acceleration is considered, the slope in the $\ln(t_{sw})$-V diagram can be directly related to the physical parameters of the underlying processes. Several processes can limit the switching speed in different voltage regimes. Then it is expected that the process with the steepest slope appears at the lowest voltage and the processes with flatter slopes appear at higher voltages. When the dissipated power is sufficiently high, temperature acceleration sets in. This onset appears in the $\ln(t_{sw})$-V diagram as a sudden decrease in switching time. As the temperature increase is power-dependent, devices with different initial resistance should show very similar characteristics in a t_{sw}-P diagram, but vary in the $\ln(t_{sw})$-V diagram.

The analysis of the published VCM SET switching kinetics data showed that their nonlinearity is mainly dominated by temperature-accelerated ion hopping. The investigated time regimes, however, typically span only over a few orders of magnitude. Thus, different processes might limit the switching speed in other regimes. The gradual RESET transition can be explained in terms of temperature-accelerated ion movement with ion drift and diffusion approaching equilibrium.

The nonlinearity in the ECM switching kinetics is governed by electric field/voltage acceleration of the underlying physical processes. The SET switching kinetics are determined by nucleation, electron-transfer and ion hopping process. For the RESET process only the electron-transfer reactions and the ion hopping process determine the switching speed.

References

[1] J. J. Yang, D. B. Strukov, D. R. Stewart, Nat. Nanotechnol. 8 (2013) 13.

[2] D. Jeong, R. Thomas, R. Katiyar, J. Scott, H. Kohlstedt, A. Petraru, C. Hwang, Rep. Prog. Phys. 75 (2012).

[3] R. Waser, R. Dittmann, G. Staikov, K. Szot, Adv. Mater. 21 (2009) 2632.

[4] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. T. Chen, M.-J. Tsai, Proc. IEEE 100 (2012) 1951.

[5] A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, R. S. Williams, Nanotechnology 22 (2011) 485203.

[6] M. Meier, C. Schindler, S. Gilles, R. Rosezin, A. Rudiger, C. Kügeler, R. Waser, IEEE Electron Device Lett. 30 (2009) 8.

[7] C. Hermes, M. Wimmer, S. Menzel, K. Fleck, G. Bruns, M. Salinga, U. Boettger, R. Bruchhaus, T. Schmitz-Kempen, M. Wuttig, R. Waser, IEEE Electron Device Lett. 32 (2011) 1116.

[8] M. N. Kozicki, M. Park, M. Mitkova, IEEE Trans. Nanotechnol. 4 (2005) 331.

[9] M.-J. Lee, C. B. Lee, D. Lee, S. R. Lee, M. Chang, J. H. Hur, Y.-B. Kim, C. -J. Kim, D. H. Seo, S. Seo, U.-I. Chung, I.-K. Yoo, K. Kim, Nat. Mater. 10 (2011) 625.

[10] T. Ninomiya, S. Muraoka, Z. Wei, R. Yasuhara, K. Katayama, T. Takagi, IEEE Electron Device Lett. 34 (2013) 762.

[11] S. Menzel, M. Salinga, U. Böttger, M. Wimmer, Advanced Functional Materials 25 (2015) 6306–6325.

[12] Y. Chen, L. Goux, S. Clima, B. Govoreanu, R. Degraeve, G. Kar, A. Fantini, G. Groeseneken, D. Wouters, M. Jurczak, IEEE Trans. Electron Devices 60 (2013) 1114.

[13] S. Yu, H.-Y. Chen, B. Gao, J. Kang, H.-S. P. Wong, Acs Nano 7 (2013) 2320.

[14] J. J. Yang, M. Zhang, J. P. Strachan, F. Miao, M. D. Pickett, R. D. Kelley, G. Medeiros-Ribeiro, R. S. Williams, Appl. Phys. Lett. 97 (2010) 232102/1.

[15] Z. Wei, Y. Kanzawa, K. Arita, Y. Katoh, K. Kawai, S. Muraoka, S. Mitani, S. Fujii, K. Katayama, M. Iijima, T. Mikawa, T. Ninomiya, R. Miyanaga, Y. Kawashima, K. Tsuji, A. Himeno, T. Okada, R. Azuma, K. Shimakawa, H. Sugaya, T. Takagi, R. Yasuhara, H. Horiba, H. Kumigashira, M. Oshima, IEEE Tech. Dig. (2008).

[16] T. Ninomiya, Z. Wei, S. Muraoka, R. Yasuhara, K. Katayama, T. Takagi, IEEE Trans. Electron Devices 60 (2013) 1384.

[17] J. Park, K. P. Biju, S. Jung, W. Lee, J. Lee, S. Kim, S. Park, J. Shin, H. Hwang, IEEE Electron Device Lett. 32 (2011) 476.

[18] F. Lentz, B. Roesgen, V. Rana, D. J. Wouters, R. Waser, IEEE Electron Device Lett. 34 (2013) 996.

[19] J. J. Yang, M. D. Pickett, X. Li, D. A. A. Ohlberg, D. R. Stewart, R. S. Williams, Nat. Nanotechnol. 3 (2008) 429.

[20] A. Sharma, M. Noman, M. Abdelmoula, M. Skowronski, J. Bain, Adv. Funct. Mater. 24 (2014) 5522–5529.

[21] R. Muenstermann, T. Menke, R. Dittmann, R. Waser, Adv. Mater. 22 (2010) 4819.

[22] N. Aslam, V. Longo, C. Rodenbuecher, F. Roozeboom, W. M. M. Kessels, K. Szot, R. Waser, S. Hoffmann-Eifert, J. Appl. Phys. 116 (2014) 64503/1.

[23] S. Lee, J. S. Lee, J.-B. Park, Y. K. Kyoung, M.-J. Lee, T. W. Noh, APM 2 (2014) 066103.

[24] J. J. Yang, F. Miao, M. D. Pickett, D. A. A. Ohlberg, D.R. Stewart, C. N. Lau, R. S. Williams, Nanotechnology 20 (2009) 215201.

[25] A. Marchewka, R. Waser, S. Menzel, 2015 International Conference On Simulation of Semiconductor Processes and Devices (SISPAD), Washington D.C, USA (2015) 297.

[26] A. Marchewka, B. Roesgen, K. Skaja, H. Du, C.-L. Jia, J. Mayer, V. Rana, R. Waser, S. Menzel, Advanced Electronic Materials (online) (2015).

[27] J. H. Hur, M.-J. Lee, C. B. Lee, Y.-B. Kim, C.-J. Kim, Phys. Rev. B 82 (2010) 155321.

[28] I. Valov, M. N. Kozicki, J. Phys. D Appl. Phys. 46 (2013) 074005.

[29] R. Waser, M. Aono, Nat. Mater. 6 (2007) 833.

[30] S. Menzel, U. Böttger, R. Waser, J. Appl. Phys. 111 (2012) 014501/1.

[31] M. Noman, W. Jiang, P. A. Salvador, M. Skowronski, J. A. Bain, Appl. Phys. A - Mater. Sci. Process. 102 (2011) 877.

[32] S. Koveshnikov, K. Matthews, K. Min, D. Gilmer, M. Sung, S. Deora, H. Li, S. Gausepohl, P. Kirsch, R. Jammy, Technical Digest - International Electron Devices Meeting, IEDM (2012) 20.4.1.

[33] D. Ielmini, F. Nardi, S. Balatti, IEEE Trans. Electron Devices 59 (2012) 2049.

[34] T. Diokh, E. Le-Roux, S.Jeannot, M. Gros-Jean, P. Candelier, J. F. Nodin, V. Jousseaume, L. Perniola, H. Grampeix, T. Cabout, E. Jalaguier, M. Guillermet, B. De Salvo, 2013 Ieee International Reliability Physics Symposium (irps) (2013) 5E.4.1.

[35] S. Yu, Y. Wu, H. Wong, Appl. Phys. Lett. 98 (2011) 103514/1.

[36] M. G. Cao, Y. S. Chen, J. R. Sun, D. S. Shang, L. F. Liu, J. F. Kang, B. G. Shen, Appl. Phys. Lett. 101 (2012) 203502.

[37] F. Alibart, L. Gao, B. D. Hoskins, D. B. Strukov, Nanotechnology 23 (2012) 75201/1.

[38] Y. Nishi, S. Menzel, K. Fleck, U. Boettger, R. Waser, IEEE Electron Device Lett. PP (2013) 1.

[39] K. Fleck, U. Böttger, R. Waser, S. Menzel, IEEE Electron Device Lett. 35 (2014) 924.

[40] S. Menzel, M. Waters, A. Marchewka, U. Böttger, R. Dittmann, R. Waser, Adv. Funct. Mater. 21 (2011) 4487.

[41] L. Goux, Y. Chen, L. Pantisano, X. Wang, G. Groeseneken, M. Jurczak, D. J. Wouters, Electrochem. Solid State Lett. 13 (2010) G54.

[42] J. H. Oh, K. C. Ryoo, S. Jung, Y. Park, B. G. Park, Jpn. J. Appl. Phys. 51 (2012) 4DD16/1.

[43] F. Nardi, S. Larentis, S. Balatti, D. Gilmer, D. Ielmini, IEEE Trans. Electron Devices 59 (2012) 2461.

[44] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, H. P. Wong, IEEE Trans. Electron Devices 58 (2011) 2729.

[45] J. H. Hur, K. M. Kim, M. Chang, S. R. Lee, D. Lee, C. B. Lee, M. J. Lee, Y. B. Kim, C. J. Kim, U. I. Chung, Nanotechnology 23 (2012) 225702/1.

[46] L. Zhao, H. Chen, S. Wu, Z. Jiang, S. Yu, T. Hou, H. P. Wong, Y. Nishi, Nanoscale 6 (2014) 5698.

[47] J. P. Strachan, A. C. Torrezan, G. Medeiros-Ribeiro, R. S. Williams, Nanotechnology 22 (2011) 505402/1.

[48] J. P. Strachan, A. C. Torrezan, F. Miao, M. D. Pickett, J. J. Yang, W. Yi, G. Medeiros-Ribeiro, R. S. Williams, IEEE Trans. Electron Devices 60 (2013) 2194.

[49] J. van den Hurk, S. Menzel, R. Waser, I. Valov, J. Phys. Chem. C 119 (2015) 18678.

[50] A. Nayak, T. Tsuruoka, K. Terabe, T. Hasegawa, M. Aono, Nanotechnology 22 (2011) 235201/1.

[51] A. Nayak, T. Tamura, T. Tsuruoka, K. Terabe, S. Hosaka, T. Hasegawa, M. Aono, J. Phys. Chem. Lett. 1 (2010) 604.

[52] S. Menzel, S. Tappertzhofen, R. Waser, I. Valov, PCCP 15 (2013) 6945.

[53] I. Valov, I. Sapezanskaia, A. Nayak, T. Tsuruoka, T. Bredow, T. Hasegawa, G. Staikov, M. Aono, R. Waser, Nat. Mater. 11 (2012) 530.

[54] G. Palma, E. Vianello, G. Molas, C. Cagli, F. Longnos, J. Guy, M. Reyboz, C. Carabasse, M. Bernard, F. Dahmani, D. Bretegnier, J. Liebault, B. De Salvo, Jpn. J. Appl. Phys. 52 (2013) UNSP 04CD02/1.

[55] J. R. Jameson, N. Gilbert, F. Koushan, J. Saenz, J. Wang, S. Hollmer, M. N. Kozicki, Appl. Phys. Lett. 99 (2011) 063506.

[56] U. Russo, D. Kamalanathan, D. Ielmini, A. L. Lacaita, M. N. Kozicki, IEEE Trans. Electron Devices 56 (2009) 1040.

[57] T. Tsuruoka, T. Hasegawa, I. Valov, R. Waser, M. Aono, AIP Adv. 3 (2013) 32114/1.

[58] L. Gao, S. B. Lee, B. Hoskins, H. K. Yoo, B. S. Kang, Appl. Phys. Lett. 103 (2013) 43503/1.

[59] T. Tsuruoka, K. Terabe, T. Hasegawa, I. Valov, R. Waser, M. Aono, Advanced Functional Materials 22 (2012) 70.

[60] S. Tappertzhofen, I. Valov, T. Tsuruoka, T. Hasegawa, R. Waser, M. Aono, ACS Nano 7 (2013) 6396.

[61] S. Tappertzhofen, R. Waser, I. Valov, ChemElectroChem 1 (2014) 1287.

[62] K. Kinoshita, ECS Trans. 69 (2015) 11.

[63] J. G. Simmons, J. Appl. Phys. 34 (1963) 1793.

[64] S. Menzel, R. Waser, Nanoscale 22 (2013) 11003.

D5 Electronic Avalanche in Narrow Gap Mott Insulators

and Non-Volatile Memories

E. Janod, B. Corraze, Julien Tranchant,
Marie-Paule Besland and L. Cario
Institut des Matériaux Jean Rouxel (IMN), Nantes, France
etienne.janod@cnrs-imn.fr

Contents

1 Introduction **2**

2 Mott insulators and Mott insulator to metal transitions **3**
2.1 Basic concepts 3
2.2 Examples of canonical Mott insulators 6
2.3 Insulator to metal transition in other correlated insulators 7

3 Resistive switching in correlated insulators **9**
3.1 Resistive switching related to temperature-controlled insulator to metal transitions 9
3.2 Valence change memories (VCM) with Mott Insulators 9
3.3 Resistive switching induced by dielectric breakdown. 10

4 Electric field induced dielectric breakdown in Mott insulators **11**
4.1 First evidence of an avalanche breakdown in AM_4Q_8 narrow gap Mott insulators 11
4.2 Electric-field-induced avalanche and dielectric breakdown : the Fröhlich model 12
4.3 Modeling of avalanche phenomena in Mott Insulators 16
4.4 Avalanche phenomena in Mott Insulators: a universal property 18

5 Non Volatile resistive switching in Mott insulators **19**
5.1 Evidence of electric field driven non-volatile Mott IMT 19
5.2 From volatile to non volatile resistive switching: control of SET and RESET 20
5.3 Electric-field-induced electronic phase separation and resistive switching 23
5.4 Towards a microscopic view of the resistive switching in Mott insulators 25

6 ReRAM devices based on avalanche breakdown in narrow gap Mott insulators **26**
6.1 Preparation of GaV_4S_8 thin active layers and GaV_4S_8 based MIM structures 26
6.2 Resistive switching in GaV_4S_8 MIM structures 26
6.3 Performances of GaV_4S_8 based ReRAM devices 27

7 Conclusion **28**

1 Introduction

The huge non-volatile memory market is led by the Flash technology, used *e.g.* in Flash SD cards and Solid State Drives. However, the limit of this technology in downscaling will hinder its development in a near future.[1] Several emerging Random Access Memories (RAM),[2] *i.e.* Phase-Change RAM (PCRAM),[3] Magnetic RAMs (MRAM),[4] and Resistive RAM (ReRAM),[5] are currently considered as interesting candidates to overcome the shortcomings of Flash memories. However, ReRAMs appear as a very appealing solution among these potential candidates, thanks to a very simple architecture and promizing memory performances.[1] ReRAMs are hence envisioned to replace the Flash technology in mass storage applications before 2020.[6] In ReRAM information storage is enabled by a non-volatile and reversible switching between two different resistance states of an active material. This resistive switching is obtained by applying short electric pulses to the active material most of the time sandwiched between two metallic electrodes. A large variety of materials are known to exhibit a reversible electric-pulse-induced resistive switching phenomenon, such as transition metal oxides Band Insulators (TiO_2, $SrTiO_3$, $SrZrO_3$ …) or copper and silver based chalcogenides.[7,8,9,10,11] So far, different mechanisms based on thermochemical or electrochemical effects have been proposed to explain the non-volatile resistive switching observed in these materials.[7,5,9] But resistive switching is also observed in Mott insulators that form a large class of materials particularly attractive in the context of memory applications.[11] They can indeed undergo various kinds of insulator to metal transitions (IMT) in response to different external perturbations like pressure, temperature, and electronic filling. These IMT are often associated with huge modifications of the electrical resistance and therefore allows generating high and low resistance states *i.e.* the two logical states (‘0’ and ‘1’) of a ReRAM device.

We will focus here on this particularly interesting class of ReRAM in which the active material is a Mott Insulator. Section 2 describes briefly the theoretical background of Mott insulators, and the different ways to break the Mott insulating state to induce insulator to metal transitions (IMT) in these systems. Most resistive switching in Mott insulators are closely related to these IMT that can be induced by electric pulses either thanks to Joule heating, or by means of electrochemical or thermochemical mechanisms. Section 3 gives an overview of resistive switching in Mott insulators and correlated insulators based on these known mechanisms first evidenced in band insulators or amorphous insulators. Conversely, a new mechanism of resistive switching was recently discovered in Mott insulators.[12,13,14] This lecture focuses particularly on this new type of resistive switching which is triggered by an electric field induced avalanche breakdown ultimately leading to a non volatile electronic phase separation at the nanoscale. Sections 4 and 5 describe, respectively, the volatile avalanche breakdown phenomenon and its non volatile consequences in Mott insulators. Section 6 displays the potential of this universal property of narrow gap Mott insulators for ReRAM applications. Finally, a classification of resistive switching mechanisms in Mott insulators is proposed, based on the types of insulator to metal transition and controlling parameters involved in the resistive switching.

2 Mott insulators and Mott insulator to metal transitions

2.1 Basic concepts

Unlike conventional band insulators and semiconductors, Mott insulators contain unpaired electrons in their ground state. However, a drastic condition is required to bring up the Mott insulating state: an electronic filling exactly equal to an integer number of unpaired electron per site.[15] According to conventional band theories, such compounds with an odd number of electrons should be metallic since their Fermi levels lie in the middle of a band, as shown in **Figure 1**. However, even in absence of disorder, many of these materials are actually insulators. The discrepancy comes from a crucial parameter incorrectly described in conventional band theories, the on-site coulombian repulsion. For simplicity, let's consider the situation of a single band system that is half-filled (*i.e.* with one electron per site). Thus, if the Coulomb repulsion (Hubbard) energy U exceeds the bandwidth W, the half-filled band splits into two sub-bands, the Lower (LHB) and Upper (UHB) Hubbard Bands (see **Figure 1**a). Thanks to the U term, a Mott-Hubbard gap E_G opens up between the LHB and the UHB if U is larger than the bandwidth W, roughly equal to $E_G \approx U - W$. The theoretical description of the Mott insulating state has been a long-standing problem[16,17,18] and only modern approaches such as the dynamical mean field theory (DMFT) have successfully predicted the whole phase diagram of this class of materials.[19,20] A salient feature of this universal k_BT/W *vs.* U/W phase diagram[19,20,21,22] is the first order (Mott) transition line which separates a metallic domain at low U/W from a Paramagnetic Mott Insulator (PMI) domain for $U/W > 1.15$,[23] as depicted in **Figure 1**a. This Mott metal-insulator transition line terminates at a second order critical endpoint at high temperature for $T_{endpoint} \approx 0.025W/k_B$.[19,22] This endpoint has an interesting fundamental consequence: the absence of crystallographic symmetry breaking across the Mott line, since one can connect continuously the PMI and metallic phases shown in **Figure 1**a through a high temperature path above the endpoint. Another major contribution of the DMFT is to predict a specific signature of electronic correlation close to the Mott IMT line : while a gap between the Lower and Upper Hubbard Bands exists on the insulating side, a quasiparticle peak develops in the gap at the Fermi energy on the metallic side (see **Figure 1**a).[19] Whereas the Mott line and the high temperature part of the phase diagram are universal, the low temperature part is material-dependent and can present various kinds of long-range (for example magnetic or orbital) orders.

Beyond the particular case of half-filling, **Figure 1**b presents a generalized phase diagram at any electronic filling, represented as x (hole or electron doping level away from half-filling) *vs.* U/W.[24] This diagram reveals that the Mott insulating state is stable only at half-filling and that doped Mott insulators are metallic. Both phase diagrams highlight the three insulator to metal transitions (IMT), represented by red arrows in **Figure 1**, that emerge from the Mott insulating state:[24]

(1) Bandwidth controlled IMT. This IMT, noted "type 1" thereafter, corresponds to the crossing of the Mott transition line (see **Figure 1**a) induced by tuning the correlation strength U/W,[24,25,26,27]This can be achieved by applying an external pressure which enhances the orbitals overlaps and increases thus the bandwidth W;

(2) Temperature controlled IMT. This IMT, noted "type 2" thereafter, is driven by temperature and also relies on the crossing of the Mott line in a narrow window of U/W around ≈ 1.15, between the red dotted line in **Figure 1**a. This IMT occurs between a *low temperature metal* and a *high temperature insulator*,[28] which strongly contrasts with the more usual transitions from a low-T insulator to a high-T metal,

(3) Filling controlled IMT.[24] This IMT, noted "type 3" thereafter, occurs when the band filling deviates from half filling (see **Figure 1**b). This may be achieved by tuning the electronic filling thanks to chemical doping.

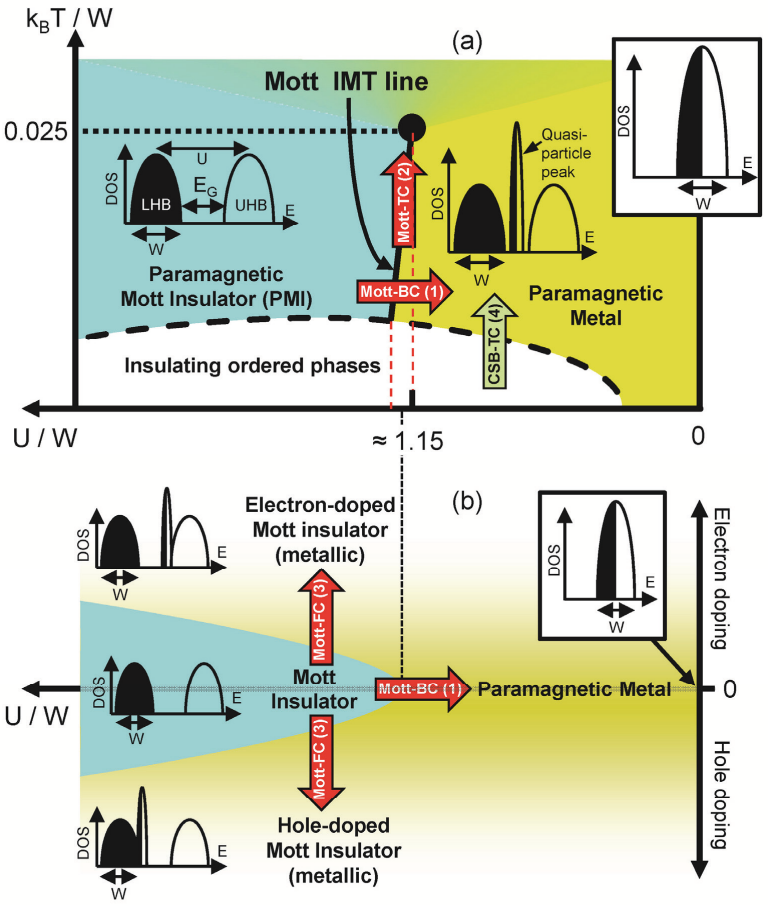


Figure 1 *: (a) Schematic phase diagram of two- and three-dimensional half-filled compounds undergoing a Mott insulator to metal transition, displayed as k_BT/W vs. U/W. T, U and W are the temperature, the Hubbard electron-electron repulsion term and the bandwidth, respectively. Typical electronic Density of States (DOS) are displayed in relevant regions of the phase diagram : in absence of electron correlation (U/W = 0), in the correlated metal domain slightly below (U/W)c = 1.15 and in the Mott insulating state for U/W > 1.15.*
(b) Diagram of doped Mott insulators, represented as electron and hole doping away from half-filling vs. correlation strength U/W, for intermediate temperature $T_{ordered\ phase} < T < T_{endpoint}$ shown in part (a).

Red arrows indicate the universal Insulator to Metal Transitions (IMT) that emerge from the Paramagnetic Mott Insulator state, *i.e.* the "type 1" Mott – Bandwith Controlled (Mott-BC) and the "type 2" Mott – Temperature Controlled (Mott-TC) transitions crossing the Mott line in half-filled compounds, as well as the "type 3" Filling-Controlled (Mott-FC) IMT. The green arrow corresponds to a non universal "type 4" Temperature-Controlled insulator to metal transition towards a long range order insulating state, associated in real systems to a Crystallographic Symmetry Breaking (CSB-TC).

These phase diagrams of correlated compounds call for several interesting remarks. **Figure 1**a indeed shows that a canonical *Paramagnetic Mott Insulator (PMI) can NOT undergo an Insulator to Metal Transition by increasing temperature*.[29] However, temperature-controlled insulator to metal transitions are possible in half-filled correlated systems if they exhibit a long-range (*e.g.* magnetic or orbital) order at low temperature, as shown by the green arrow in **Figure 1**a. In real systems, such Temperature Controlled IMT involve a symmetry breaking, which is in general a crystallographic symmetry breaking. These insulator to metal transitions, noted "type 4" thereafter, strongly differ from the three Mott IMT discussed above which occur *without* crystallographic symmetry breaking. In the compounds showing a "type 4" temperature controlled IMT involving a crystallographic symmetry breaking, the driving force behind the IMT is not related only to the *U vs. W* competition, but necessarily includes an additional mechanism.

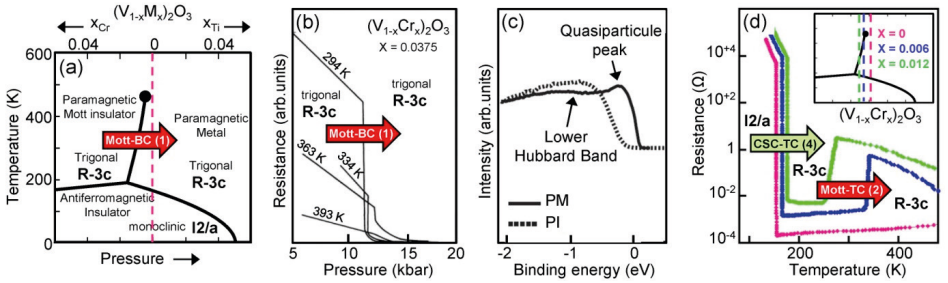


Figure 2 *: (a) Phase diagram of $(V_{1-x}M_x)_2O_3$, with M = Cr and Ti. In this system, changing the V/M ratio by 1% is equivalent to applying an external pressure of ≈ 4 kbar.[35]*
(b) Resistance versus pressure across the Mott insulator to metal transition in $(V_{0.9625}Cr_{0.0375})_2O_3$. Adapted with permission from Ref. [36]. Copyright © 1970, American Physical Society.
(c) Photoemission spectra taken at 300 K (in the Paramagnetic Mott Insulator state, PI) and 200 K (Paramagnetic Metal state, PM) from the (001) surface of $(V_{0.989}Cr_{0.011})_2O_3$. The black arrows highlight the Lower Hubbard Band in the PI state, on top of which a quasiparticle peak appears in the PM state. Adapted with permission from Ref. [37]. 2009, American Physical Society.
(d) Resistivity vs. temperature in pure V_2O_3 and $(V_{1-x}Cr_x)_2O_3$ with x=0.006 and 0.012. Adapted with permission from Ref. [28]. Copyright © 1980, American Physical Society.

In addition, the phase diagram *doping versus U/W* (**Figure 1**b) shows the existence of a critical doping $x_c \neq 0$ necessary to induce a Mott insulator to metal transition. This critical doping increases with the correlation strength *U/W* and thus with the Mott-Hubbard gap. Experimentally, this trend is well illustrated by the $RTiO_3$ (R = La, Pr, Nd, Sm, Y) Mott insulators, with a critical doping increasing from $x_c \approx 0.03$ for $LaTiO_3$ (Mott-Hubbard gap $E_G \approx 0.1$ eV) to $x_c \approx 0.35$ for $YTiO_3$ ($E_G \approx 0.45$ eV).[30] Theoretically, the issue of a finite doping necessary to achieve an insulator to metal transition both in band and Mott insulators has been

first discussed on the basis of the "Mott criterion"[31] (see lecture of M. Wuttig on Electron Transport - Disorder and Correlation for more details). A more recent theoretical development specific to Mott insulators proposes the existence of a finite x_c which scales with $\sqrt{U/W - (U/W)_C}$, in good agreement with the behavior observed in titanates.[32]

2.2 Examples of canonical Mott insulators

The phase diagrams shown in **Figure 2** are purely theoretical and an important issue is to establish their relevance in real compounds. The most famous "canonical" Mott insulator is probably the oxide compound $(V_{1-x}Cr_x)_2O_3$. Its phase diagram[33,34,35] shown in **Figure 2**a indeed compares very well with theoretical predictions. It contains a Mott IMT line ending around 450 K and separating a Mott Insulating phase from a metallic phase. **Figure 2**b shows that applying a moderate pressure increases the bandwidth[25] and induces a type 1 (Mott) bandwidth-controlled IMT in $(V_{0.9625}Cr_{0.0375})_2O_3$.[36] Moreover, despite the strong decrease of unit cell volume at the IMT indicating a first order transition, the IMT occurs between two R-3c phases, *i.e.* without any crystallographic symmetry breaking.[35] Also, the observation of a quasiparticle peak above the Lower Hubbard Band, shown in **Figure 2**c, confirms the correlated nature of the metallic state in pure V_2O_3.[37] Finally, pure and Cr-substituted V_2O_3 display an antiferromagnetic insulating (AFI) phase at low temperature ; in pure V_2O_3, an IMT occurs between the AFI (space group I2/*a*) and the metallic phase (space group R-3c) at ≈ 165 K.[38] According to the classification proposed in Section 2.1, this transition does not correspond to a Mott transition. It corresponds to a "type 4" IMT, *i.e.* a transition associated with a crystallographic symmetry breaking and driven by an additional mechanism which is magnetic ordering in this case. **Figure 2**d shows that two successive transitions (type 4, AFI ➔ metal and type 2, metal➔ PMI) appear in a narrow V/Cr substitution level around 1%.[28] The type 2 low temperature metal to high temperature paramagnetic insulator is expected from the theoretical phase diagram of **Figure 1**a, as the Mott IMT line is not vertical but slightly tilted.[19,21,22] All these features indicate that the V_2O_3 system is a prototypical Mott insulator.

Beyond the V_2O_3 system, a few other canonical Mott insulators have been identified, such as the 2D molecular family κ-(BEDT-TTF)$_2$X [39] or the chalcogenide system $NiS_{2-x}Se_x$.[40] Recently another series of chalcogenides, the AM_4Q_8 compounds (A=Ga, Ge; M= V, Nb, Ta, Mo; Q=S, Se, Te), has emerged as a potential new example of canonical Mott insulator. These compounds exhibit a lacunar spinel structure, in which the electronic sites correspond to the tetrahedral transition metal clusters M_4 shown in the inset of **Figure 3**a.[41] In GaM_4Q_8 compounds, each M_4 cluster contain one unpaired electron among seven (M = V, Nb, Ta) or eleven (M=Mo) *d* electrons.[42] These compounds own a narrow gap of 0.1-0.3 eV, which can be tuned by chemical substitution.[43] At ambient pressure, all AM_4Q_8 compounds display two important characteristics of canonical Mott insulators : they are paramagnetic insulators above 55K [42,44,45] and do *not* exhibit any temperature-controlled IMT up to 800 K, as shown in **Figure 3**a. Moreover, these compounds exhibit a bandwidth-controlled IMT (type 1). $GaTa_4Se_8$ and $GaNb_4Q_8$ (Q=S, Se) undergo indeed an insulator to metal transition under pressure, with superconductivity at $T_C \approx$ 2-7 K in the pressurized metallic state above 11 GPa.[46,47] Recent studies of transport properties under pressure in $GaTa_4Se_8$, shown in **Figure 3**b, prove that this pressure-induced (bandwidth-controlled) IMT is of first order with an hysteresis, as expected from LDA+DMFT calculations.[48,49] Moreover, the optical conductivity shown on **Figure 3**c reveals the signature of a quasi-particle peak in the pressurized metallic phase of $GaTa_4Se_8$.[50] Another interesting feature of AM_4Q_8 is that they undergo filling-controlled IMT (type 3) when doped on the A site or on the M site.[51] All these results demonstrate that the AM_4Q_8 compounds display the expected characteristics of a canonical Mott insulator.

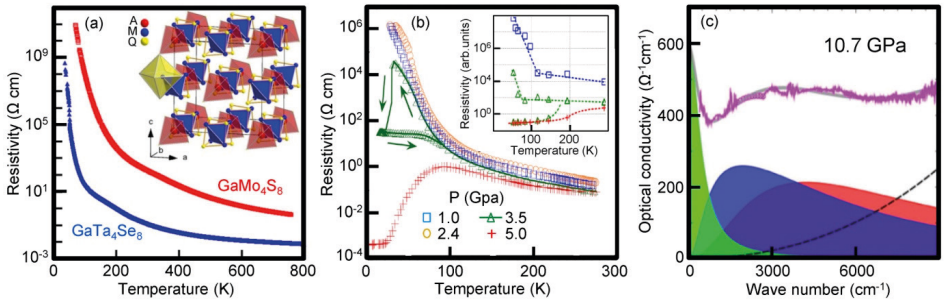


Figure 3 *: (a) Resistivity vs. temperature up to 800 K in two representative AM_4Q_8 compounds, $GaMo_4S_8$ and $GaTa_4Se_8$. Inset : crystallographic structure of AM_4Q_8 (A = Ga, Ge; M = V, Nb, Ta, Mo; Q = S, Se) compounds, highlighting the M_4 tetrahedral clusters.*
(b) Resistivity vs. temperature (4 K ≤ T ≤ 300 K) at different pressures in $GaTa_4Se_8$ in the PMI (1 and 2.4 GPa) and metallic (5 GPa) states. The "bistability" of resistivity at 3.5 GPa is a clear indication of the phase coexistence close to the Mott IMT line. Inset: LDA + DMFT results for the resistivity as a function of the temperature. The red crosses, blue squares, and green triangles correspond to the metal, insulator and coexistent solutions, respectively. The lines are a guide for the eyes. Reproduced from Ref. [48].
(c) Optical conductivity vs. wave number in $GaTa_4Se_8$ in the metallic state appearing beyond the Mott line under pressure (10.7 GPa). The low energy contribution corresponds to the quasiparticle peak, a typical signature of electronic correlation. Reproduced from Ref. [50].

2.3 Insulator to metal transition in other correlated insulators

Beyond the examples of canonical Mott insulators and Mott IMT, many other half-filled insulators display Temperature-Controlled IMT potentially interesting for memory applications. Most of these IMT are clearly not of the Mott type 1, 2 or 3 discussed above, but belongs to the type 4 IMT since they are associated with crystallographic symmetry breakings. **Figure 4** gathers several examples of such "type 4" insulator to metal transitions, which include IMT in Ca_2RuO_4 (T_{IMT} = 357 K),[52] VO_2 (T_{IMT} = 340 K),[53] NbO_2 (T_{IMT} = 1070 K)[54] and $ANiO_3$ perovskites.[55,56] As illustrated by the representative example of VO_2 shown in **Figure 5**, the temperature-pressure phase diagram of these half-filled insulators contains, unlike canonical systems (see **Figure 1**a and **Figure 2**a), an IMT line separating a low-*T* insulating phase from a high-*T* metallic phase of different crystallographic symmetry.[57,58]

Finally it is worth mentioning that temperature controlled IMT can also happen in non-half-filled correlated systems. In mixed valence systems, insulator to metal transition may indeed go along with a charge ordering transition, as observed *e.g.* in 2D molecular systems [59] and in transition metal oxides (see Ref. [60] for a short review). Such IMT are always accompanied by crystallographic distortions to low symmetry in the charge ordered insulating phase at low temperature. They are thus related with the temperature-controlled IMT (type 4) of half-filled systems discussed above. A prominent example is the Verwey transition occurring at 122 K in the magnetite Fe_3O_4.[61]

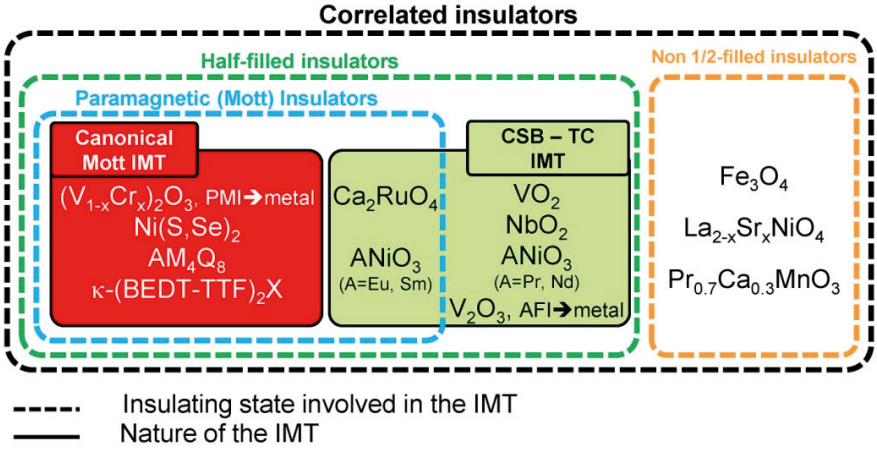


Figure 4 *: Classification of insulator to metal transitions (IMT) occuring in various correlated insulators of interest for resistive switching effects. Unlike canonical (Mott) IMT which result only from a competition between U and W (IMT in* $(V_{1-x}Cr_x)_2O_3$*,* $Ni(S,Se)_2$*,* AM_4Q_8 *and* κ*(BEDT-TTF)*$_2$*X), Temperature-Controlled IMT associated with a Crystallographic Symmetry Breaking (CSB-TC) are driven by another mechanism : Jahn-Teller effect at* T_{IMT}*=357 K in* Ca_2RuO_4 *[52], a Peierls-Mott instability at* T_{IMT} *= 340 K in* VO_2 *[53] and at* T_{IMT} *= 1070K in* NbO_2 *[54], magnetic ordering at* T_{IMT}*=165K in pure* V_2O_3 *[38] or a complex site-selective transition in* $ANiO_3$ *perovskites [55]. The compounds gathered on the right hand side are typical examples of non half-filled systems where the insulating state results from a charge ordering.*

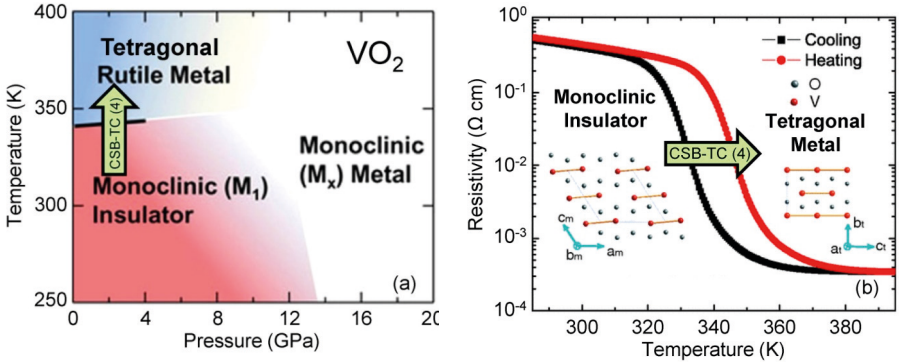


Figure 5 *: (a) Temperature-pressure phase diagram of* VO_2*. Adapted with permission from Ref. [57], Copyright 2014, American Institute of Physics. (b) Resistivity vs. temperature at the "type 4" insulator to metal transition (IMT) in* VO_2*. This IMT is associated with a Crystallographic Symmetry Breaking between the monoclinic low-T and the tetragonal high-T phases. Reproduced with permission from Ref. [58], Copyright 2013,American Physical Society.*

3 Resistive switching in correlated insulators

3.1 Resistive switching related to temperature-controlled insulator to metal transitions

In Section 2, two different classes of thermally driven IMT were introduced, occurring with (type 4 IMT) or without (type 2 IMT) crystallographic symmetry breaking. For both types of IMT, temperature can be used as a tuning parameter triggering a resistive switching. Indeed, the application of an electric field at $T < T_{IMT}$ can lead to Joule self-heating and therefore to a strong modification of resistance if the sample temperature exceeds T_{IMT}.

Such a thermal mechanism is at play in correlated metal in the close vicinity of the Mott line, as recently confirmed by a DMFT theoretical study.[62] In compounds such as $(V_{1-x}Cr_x)_2O_3$ ($x \approx 0.01$)[63] and $NiS_{2-x}Se_x$ ($x \approx 0.45$),[64,65]a volatile resistive switching under electric field indeed occurs due to Joule heating effects, between a low T metallic phase and a high T paramagnetic Mott insulator phase (see **Figure 2**a). A more recent work on $GaTa_4Se_8$ under pressure also underlines the important role of Joule heating near the Mott IMT line.[48] This switching are related to a Mott type 2 IMT and leads to an *increase* of resistance during the pulse.

Also, this thermal mechanism convincingly explains the switchings observed in the compounds displaying a type 4 IMT, such as in VO_2,[66,67] NbO_2,[68] Ca_2RuO_4,[69,5] in pure V_2O_3 below the AFI - metal transition temperature [70,71,72] and in magnetite Fe_3O_4.[73,74] These thermally-induced switchings are essentially volatile (*i.e.* low resistance state is maintained only under electric field) and appears above a threshold voltage corresponding to a Joule heating threshold. The materials showing such a *volatile threshold switching* behavior [75] can be used as selectors in Resistive Random Access Memory (ReRAM) crossbar arrays, in order to suppress the undesired sneak currents (see Ref. [76] for a general introduction on this concept). However, non-volatile *resistive switching* (*i.e.* the low resistance state remains even after the end of electric pulses) can be also achieved in these correlated insulators by fine tuning the working temperature within the hysteresis domain of the first order IMT, as demonstrated in VO_2.[77,78] However this compound was barely studied in the context of ReRAM applications.

3.2 Valence change memories (VCM) with Mott Insulators

Resistive switching based on valence change is one of the most known mechanisms for ReRAM, and has been the focus of many reviews. [7,5,8,9,10,11] In non stoichiometric transition metal oxides like $SrTiO_{3-x}$, TiO_{2-x}, HfO_{2-x}, Ta_2O_{5-x} the migration of oxygen vacancies under electric field along grain boundaries or dislocations induces a valence change of the cations in the vicinity of theses defects.

At the local scale, a transition occurs between a band insulator involving empty *d*-orbitals (d^0) with cations in their high valence state (*e.g.* Ti^{4+}) to a metallic state (degenerated doped insulator) involving partially filled *d*-orbitals ($d^{+\delta}$) with cations in a lower valence state (*e.g.* Ti^{3+}). In oxide like $SrTiO_{3-x}$, TiO_{2-x}, HfO_{2-x}, Ta_2O_{5-x}, this phenomenon leads to a reversible bipolar resistive switching[79] by the formation/destruction of a metallic filamentary path between the electrodes. [9,5,80] Alternatively, the electro-migration phenomenon can also occur close to the metallic electrode/insulator oxide interface and lead to a bipolar resistive switching by modification of a Schottky barrier. [5,8] This interface type VCM was for example observed for $SrRuO_3/SrTi_{0.99}Nb_{0.01}O_3$/Ag junction. [8,81]

As discussed in Section 2.1, filling controlled insulator to metal transition can also occur in Mott insulators. In oxide Mott insulators this type of IMT is easily achieved by tuning the oxygen content.[24] For this reason, non stoichiometric oxide Mott insulators can exhibit both filamentary and interfacial VCM type resistive switching. Interfacial VCM type resistive switching was observed for various Mott or correlated transition metal oxides such as La_2CuO_4, [8] $Pr_{0.7}Ca_{0.3}MnO_3$, [82,83] and $YBa_2Cu_3O_{7-x}$. [84] On the other hand, filamentary VCM type resistive switching was reported in many transition metal oxide Mott insulators. [11,85] This type of resistive switching was observed for example in NiO,[86,87] CuO, [88,89] and CoO, [90] and proposed in Fe_2O_3, [91] and MnO_x. [92] The most studied system is by far NiO. In this compound, many studies have revealed that the resistive switching is related to the creation of metallic Ni filaments by a thermally assisted ionic migration process while the destruction of these filaments occurs due to Joule heating. As a consequence unipolar resistive switching[79] was mainly reported for NiO. [10,86,87] In the same way, resistive switching in CuO films was associated to the formation and destruction of conducting filaments made of a reduced phase, namely Cu_2O. [88] Conversely, resistive switching in CoO films was proposed to be related to the formation of an oxidized phase Co_3O_4. [92]

3.3 Resistive switching induced by dielectric breakdown.

As discussed in the previous sections, most of the resistive transitions observed in Mott and correlated insulators can be explained by Joule heating driven phase transition leading to an IMT, or by a filling controlled IMT induced by ionic migration. However several experimental reports of resistive switching in Mott insulators or correlated systems cannot be explained by these mechanisms. This is the case of the volatile resistive switching reported in the quasi-one-dimensional Mott insulators Sr_2CuO_3 and $SrCuO_2$ by Taguchi *et al.* [93] or in the insulating charge-ordered state of $La_{2-x}Sr_xNiO_4$. [94] A so called dielectric breakdown occurs for these compounds above a threshold field of the order of 10^2-10^4 V/cm. Similar phenomena were also reported for the family of chalcogenide Mott insulators AM_4Q_8 (A = Ga, Ge; M = V, Nb, Ta, Mo; Q = S, Se), [12,13] or for the molecular Mott insulators K-TCNQ, [95] and κ-$(BEDT\text{-}TTF)_2Cu[N(CN)_2]Br$. [96] In that context, many theoretical works were recently devoted to dielectric breakdown caused by strong electric fields in Mott insulators. These studies have mainly focused on the Zener breakdown for Mott insulators [97,98,99,100]. For instance, calculations were performed in 1D Hubbard chains using exact diagonalization, [97] and time-dependent density matrix renormalization group, [98] or in the limit of large dimensions using dynamical mean field theory. [99] All these theoretical studies have predicted non-linear behavior in the current-voltage characteristics, and the existence of a threshold field (E_{th}) beyond which a field induced metal appears. This dielectric breakdown should occur when the electric field is such that it bends the Hubbard bands by the gap energy E_G within the length ξ of the order of to the unit cell. Hence, the Zener breakdown is predicted to occur for strength of the electric field $E_{th} \sim E_G/\xi$ of the order of 10^6-10^7 V/cm. [97] This is at least two orders of magnitude larger than the values observed experimentally. [93,94,13,96] As a consequence, volatile resistive switching in Mott insulators cannot be explained by a Zener breakdown scenario. Alternatively, recent studies on the AM_4Q_8 Mott insulators support that the dielectric breakdown originates from an electric field induced electronic avalanche phenomenon. The following sections will describe in more detail the experimental evidences, theoretical modeling and ReRAM applications of this universal property of Mott Insulators.

4 Electric field induced dielectric breakdown in Mott insulators

4.1 First evidence of an avalanche breakdown in AM_4Q_8 narrow gap Mott insulators

AM_4Q_8 (A = Ga, Ge ; M = V, Nb, Ta, Mo; Q = S, Se) Mott insulators are very sensitive to electric pulses. [12,13,101,102] When an electric field pulse exceeding a threshold field (E_{th}) of a few kV/cm is applied to these compounds they undergo a sudden decrease of their resistance. As an example **Figure 6** shows the typical time evolution of the intensity *I(t)* and of the voltage $V_{sample}(t)$ across a GaV_4S_8 crystal during the application of a series of short voltage pulses to a circuit composed of the crystal connected in series with a load resistance (sketched in Figure 6a). [103] An abrupt increase of the intensity and a lowering of the voltage across the sample is observed for applied voltages that exceed the threshold voltage V_{th} (or more precisely the threshold field $E_{th} = V_{th}/d$ with d the inter-electrodes distance) shown as red dotted line in **Figure 6**b. These transitions correspond to volatile resistive switchings from a high to a low resistance state, since resistance returns to its initial value after the electric pulse terminates. It is worth noting that these transitions cannot be explained by a temperature controlled IMT (described as type 2 IMT in Section 3.1) since AM_4Q_8 compounds do not present any IMT in temperature (see Figure 3a). Moreover simple estimates using the energy release during the pulse and the activated temperature dependence of the resistivity show that Joule heating cannot account for the abrupt resistive switching. [13] Figure 6 shows that the resistive switching occurs only above a threshold electric field E_{th} (≈ 7 kV/cm for GaV_4S_8) and after a time t_{delay} which decreases as the voltage across the sample increases. The sample voltage V_{sample} after the resistive switching event always lies on the same value $V_{th} \approx 12$ V (or $E_{th} \approx 7$ kV/cm) that also corresponds to the lower voltage that can induce a resistive switch in DC measurements. The AM_4Q_8 compounds exhibit therefore a very specific current-voltage characteristics with two branches. The first one corresponds to the non transited state and follows the Ohm's law. The second branch, which is almost vertical and lies at the threshold field, corresponds to the "transited" state (see red dotted line in **Figure 6**c). All AM_4Q_8 compounds exhibit the same type of *I(V)* characteristic with threshold electric field in the 1-10 kV/cm range.[13] The magnitude of the threshold field in AM_4Q_8 Mott insulators as well as their *I(V)* characteristics compare well with the threshold field values and *I(V)* characteristics observed for avalanche breakdowns in narrow gap semiconductors.[104] For this reason it was proposed that the resistive switching observed in the Mott Insulators AM_4Q_8 originates from an avalanche breakdown phenomenon.[105] In semiconductors the avalanche threshold field varies as a power law of the band gap and follows the universal law $E_{th} \propto E_G^{2.5}$.[106,107] **Figure 6**d reveals that AM_4Q_8 compounds have a similar variation of the threshold field as a function of the Mott-Hubbard gap.[105] This power law behavior provides a first evidence that supports the avalanche breakdown scenario in these Mott insulators.

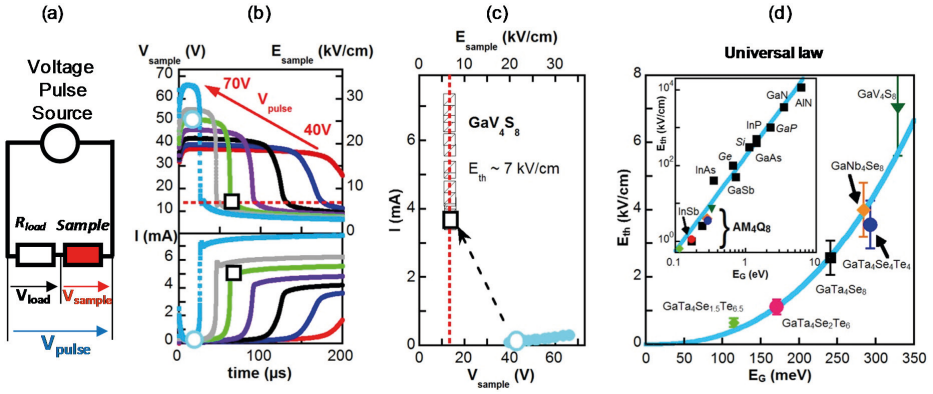


Figure 6. *(a) Example of circuit used for measurement. (b) Time dependence of the voltage and intensity across a GaV_4S_8 single crystal during 200 µs pulses for several voltages applied to the circuit. Above a threshold voltage of ≈ 12V (equivalent to 7 kV/cm), a resistive switching occurs after a time t_{delay} which decreases when the sample voltage (electric field) increases. All the transitions observed during the pulses are volatile, i.e. the resistance is the same before and after the electric pulse. (c) Current-voltage characteristics measured during the pulses, before (blue circles) and after (open squares) the volatile transition (see corresponding symbols in fig. 6(b)). (d) Dependence of E_{th} in Mott insulators and semiconductors. Threshold electric field (inducing avalanche breakdown) as a function of the Mott gap E_G for various AM_4Q_8 compounds. The solid blue curve corresponds to a power law dependence $E_{th} \propto E_G^{2.5}$. Inset: comparison of the threshold fields versus gap dependence for the AM_4Q_8 compounds and for classical semiconductors. The solid blue line displays the universal law $E_{th}[kV/cm] = 173 (E_G[eV])^{2.5}$ observed for semiconductors. Reproduced with permission.from Ref.[105]· Copyright 2013, Macmillan Publishers Limited.*

4.2 Electric-field-induced avalanche and dielectric breakdown : the Fröhlich model

In classical semiconductors such as Si, Ge or GaAs, avalanche breakdown is the consequence of an impact ionization: electrons in the conduction band, accelerated by an electric field, can gain enough energy to induce an impact ionization. This process generates electron-holes pairs and promotes new electrons in the conduction band. The repetition of this process leads to a free-carriers multiplication if the average ionization impact rate exceeds the electron-hole recombination rate. A good illustration of such an avalanche breakdown in GaAs is provided in Ref.[108].

The foundations of current theories of electrical breakdown driven by electric fields in classical semiconductors were laid more than 70 years ago by Fröhlich and Seitz [109,110,111,112,113]. These pioneering works lead to distinguish between two different regimes. A first regime, corresponding to a "clean" limit, occurs preferentially at low temperature in ultrapure semiconductors with long mean free paths: in this case, the avalanche process is initiated by the tiny number of electrons available, which gain energy independently to each others. In this regime, the electric-field-induced energy gain in the conduction band is only limited by the

electron-phonon scattering. As e^- - ph scattering raises with temperature, the threshold electric field E_{th} also increases with T in this regime.

Alternatively, a "dirty" regime may appear either at higher temperature where electron-electron scattering becomes important and in the realistic situation where defects induce discrete energy levels in the gap. Fröhlich considered a specific density of states of defects levels described in **Figure 7**.a. The trapped electrons have a ground state ε_0 located at ε_G below the conduction band, and several localized excited energy levels distributed on a typical energy width $\Delta\varepsilon$ and located just below the conduction band. Unlike the electrons in the conduction band, the trapped electrons cannot be accelerated by an electric field because these levels are not continuously distributed. However if their density is high enough they give rise to an efficient scattering with the conduction electrons.

The theory of the dielectric breakdown in the dirty limit is based on the thermal balance of the electronic system, which consists of the energy transfer rates from the electric field to the conduction electrons (gain P_{in}) and from the electrons localized in shallow levels to the lattice (loss P_{out}), as shown **Figure 7**.c. Under electric field, the electronic temperature T_e rises above the lattice-bath temperature T_0 until the two energy transfer rates (gain and loss) equilibrate (see **Figure 7**.d). More importantly, above a threshold field E_{th} the system becomes unstable as the energy rates gained and lost by the electrons can no longer be equilibrated. As a consequence the electronic temperature raises drastically which induces the electrical breakdown. In contrast to the classic avalanche breakdown in the clean limit, due to a few independent high energy electrons, this mechanism is a collective phenomenon where the energy increase is shared by all electrons.[109] The main prediction of this model is that the threshold electric field follows a thermally activated behavior :

$$E_{th} \propto \exp\left(\frac{\Delta\varepsilon}{4kT}\right) \quad (1)$$

Figure 8.b displays the temperature dependence of the threshold field E_{th} obtained for several $GaMo_4S_8$ single crystals. All samples give similar results with two temperature regimes. The low temperature regime is characterized by a saturation of E_{th} with a maximum value of 110 kV/cm at 74K, while at high temperature (above 125K) the threshold field decreases exponentially, following Equation (1) with $\Delta\varepsilon = 230$ meV. This behavior points to a remarkable agreement with several key predictions of the theory: existence of a threshold electric field, of two temperature regimes and of an activated dependence of E_{th} at high temperature. The same type of temperature dependence of the threshold field was measured for two other compounds of the AM_4Q_8 family, namely GaV_4S_8 and $GaTa_4Se_8$ (see **Figure 8**.b). For these compounds also, E_{th} exhibits a thermal activation law and a fit with Equation (1) shows that $\Delta\varepsilon$ varies from 114 meV in GaV_4S_8 to 112 meV in $GaTa_4S_8$. This experimental observation provides another strong indication that the resistive switching in AM_4Q_8 compounds originates from the creation of hot electrons under electric field.

Beyond the classical semiconductors, these results thus provide clear evidence that the electronic avalanche process can also can appear in Mott insulator. However, there are deep differences between the avalanche process in Mott insulator and in semiconductors:

- unlike semiconductors, slightly doped Mott insulators provide an almost perfect realization of the hypothetical density of state proposed by Fröhlich. The seminal theoretical works of Eskes, Meinders and Sawatzky on doped Mott insulators have indeed demonstrated that the "defects" levels associated with tiny charge doping are located right below the Upper Hubbard Band (see Ref. [114,115] for details). This issue was verified experimentally using various spectroscopic techniques, including convincing Scanning Tunneling Microscopy/Spectroscopy data shown in **Figure 7**.c. In

classical semiconductors, the position of defects levels in the gap vary according to the nature of the impurities or defects, and are not necessarily contiguous to the conduction band. From this viewpoint, "real" Mott insulators (*i.e.* compounds with a weak non-intentional electronic doping due *e.g.* to non-stoichiometry) are the ideal candidates to check Fröhlich's predictions.

- in spite of a common origin of the electronic avalanche in classical semiconductor and in Mott insulators, one can expect deep differences in the *consequences* of the avalanche in these two classes of compounds. Insofar the increase of electrical current is controlled (*e.g.* through an external current compliance), the avalanche phenomenon is purely volatile in classical semiconductors : the "conducting" state induced by the free-carriers multiplication disappears once the electric field is turned off. The deep reasons are (i) that the very existence oft the valence (VB) and conduction (CB) bands are not modified by the massive presence of hot electrons in the CB during the avalanche process in classical semiconductors; (ii) moreover, there are no states in competition (*i.e.* states with a close energy state) with the semiconducting state in this class of compounds. Conversely, we will see in Section 5 that the multiplication of free carriers induced by the electronic avalanche can lead to a non-volatile and reversible effect in Mott insulators. From a theoretical viewpoint, the exact relationship connecting the massive creation of free carriers to the formation of a metallic state in Mott insulators is still an open question. However several points can be put forward. First the nature of the Mott insulating state strongly differs from the semiconducting state: unlike semiconductors, the Mott state involves unpaired electrons at T=0K and is intimately related with the concept of half-filling (all the sites are occupied by a single localized electron). The Hubbard energy U that separates the Lower (LHB) and Upper (UHB) Hubbard Bands corresponds to the typical energy associated with the double occupancy of some sites. Therefore, unlike the valence and conduction bands of classical semiconductors, the LHB and UHB are NOT rigid bands and are prone to be deeply modified when massive electronic excitations (*i.e.* for a drastic departure from the half-filled state with one localized electron per site) occur. Another major difference appears between semiconductors and Mott insulators: in the latter, a competing state exists very close in energy, the correlated metal state (see Section 2 and phase diagram shown in **Figure 1**). The difference in energy between the Mott insulator and the correlated metal is all the more small that the Mott insulator is close to the critical $(U/W)_C$, *i.e.* that the system is a narrow gap Mott insulator. In this context, a possible scenario is that, after the massive excitation linked to the avalanche, the system can explore a wide energy landscape and may relax into the "correlated metal" state, which is metastable but very close in energy compared to the Mott insulating state. This idea is very similar to the standard concepts used in the field of photoinduced phase transitions.[116]

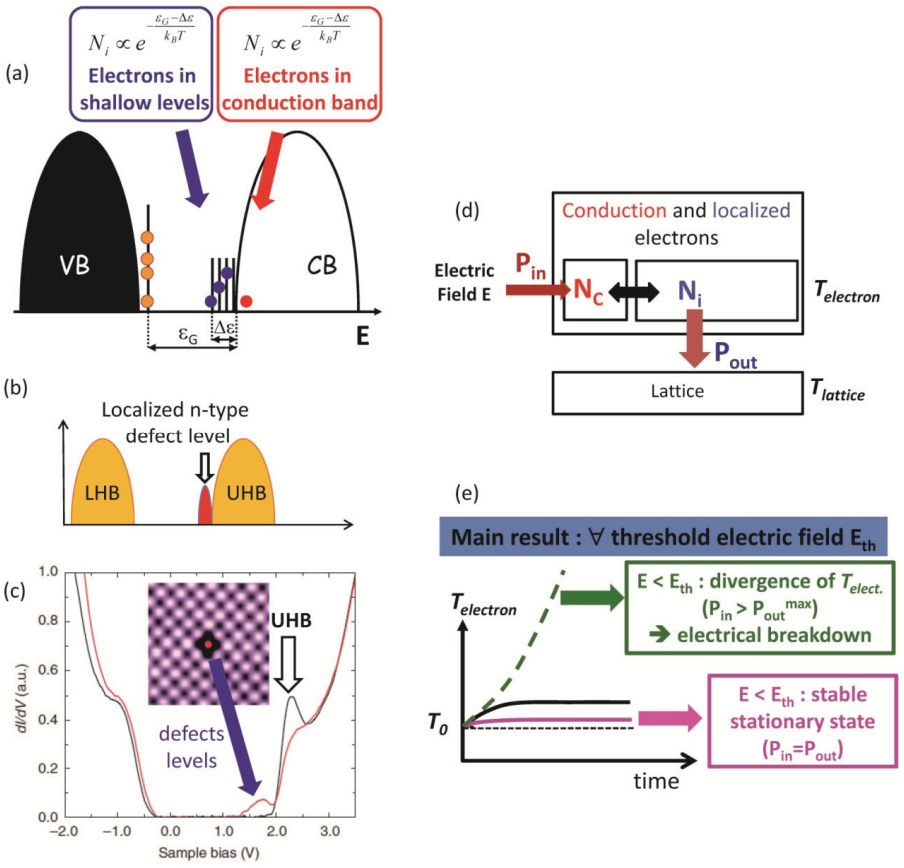


Figure 7 *: Schematic description of the main ingredients involved in the dielectric breakdown in the dirty limit according to the Fröhlich model. (a) The density of states displays a gap ε_G between the fundamental impurity level and the conduction band, with discrete impurity levels spread on the energy width $\Delta\varepsilon$. Some electrons in the impurity levels (N_i) and in the conduction band (N_c) are indicated by the blue and red points. (b) typical DOS in a slightly n-doped Mott-Hubbard insulator, according to Ref.[114,115]. An additional peak related to localized n-type defects level appears contiguous to the Upper Hubbard Band. (c) Experimental evidence of such defects level unraveled by STM/STS experiments performed close (red curve) and far (black curve) from a missing Cl defect in $Ca_2CuO_2Cl_2$. From Ref.[117]. (d) Thermal balance of the electronic system. Electrons have a temperature distinct from the lattice temperature. P_{in} and P_{out} are the heating and cooling powers sustained by the electrons. The thermalization between electrons, the heating by the electric field and the cooling through the lattice are described. (e) Time evolution of the electronic temperature under several values of the electric field. The dielectric breakdown regime appears above a threshold electric field E_{th}.*

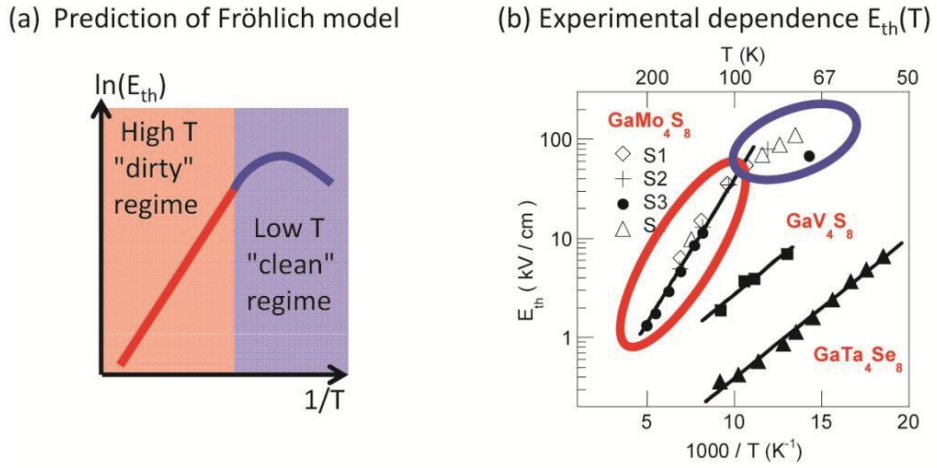


***Figure 8** : evolution of the threshold electric field E_{th} with temperature. (a) theoretical predictions of Fröhlich in the dirty (high temperature) and clean (low temperature) regimes. (b) Temperature dependence of the threshold field Eth for three compounds of the AM4Q8 family. In the case of $GaMo_4S_8$, four samples (S1-S4) have been measured. The black lines are the fit with Equation (1).*

4.3 Modeling of avalanche phenomena in Mott Insulators

Avalanche breakdown in semiconductors is related to an impact ionization process: some electrons accelerated by an electric field can promote by direct impact other electrons from the valence band to the conduction band, hence creating electron-hole pairs.

In the same way, avalanche breakdown in Mott insulators could result in the massive creation of doublons (*i.e.* doubly occupied sites) and holes at the local scale, and hence break locally the Mott insulating (MI) state into a correlated metallic (CM) state. The volatile resistive switching was therefore modeled by implementing a resistor network made of an array of cells (**Figure 9**b) which represents a small portion of the crystal that may be of a few nanometers. [105,118] Each cell is either in MI or in CM state, and its resistance is either in high or low resistance state, respectively R_{MI} or R_{CM}. The transition between both states was modeled using the energy landscape presented in **Figure 9**a. The CM state has a higher energy E_{CM} than the MI state, since the compounds are generally in the Mott insulating state and the correlated metal state is metastable. The application of an electric field increases the energy level of the MI state, and thus lowers the difference in energy between both states. In this model the MI → CM transition is mainly dependant on the electric field:

$$P_{MI \rightarrow CM} = \nu e^{-\frac{E_B - q|\Delta V|}{kT}} \qquad \text{(Eq.1)}$$

(ν is an attempt rate, q is the charge, T is the temperature and ΔV is the local voltage drop for the considered cell)

while the CM → MI transition is a thermally activated relaxation from a metastable state :

$$P_{CM \rightarrow MI} = \nu e^{-\frac{E_B - E_M}{kT}} \quad \text{(Eq. 2)}$$

This model reproduces the experimental phenomenology of the RS (*i.e.* time evolution of current and voltage and *I(V)* characteristic) and provides a microscopic view of the transition. [105,118,119] Under an applied electric field, insulating sites transform into metallic at a rate given by Eq.1. If the transformation rate overcomes the relaxation one of Eq.2, then metallic sites accumulate with time (regime depicted in yellow in **Figure 9**d) in the material. This process continues until a critical density of CM regions sets off an avalanche-like process, which ends in the formation of a conductive path connecting the electrodes (regime in green in **Figure 9**d). A typical filament is presented in **Figure 9**e, just after its creation which leads to a resistive switching. After percolation, the number of metallic sites still goes on increasing (regime represented in pink in **Figure 9**d), although at a lower rate, as long as the electric field is applied. In these three different regimes, the rate of accumulation of metallic sites accelerates when the applied electric field increases. As a consequence for higher voltage the slope for the creation of metallic sites is steeper (yellow region in **Figure 9**d) and the time for the creation of the filament is shorter (green region in **Figure 9**d). It explains the decrease of delay time after which the transition occurs *vs.* the applied voltage as found experimentally (see **Figure 9**a). [118,119] Finally, calculations combining the energy landscape model with a thermal model confirm that the onset of the resistive transition is solely driven by a purely electronic transition, while Joule heating occurs once the metallic filament is created and the current starts to raise in the circuit. [119]

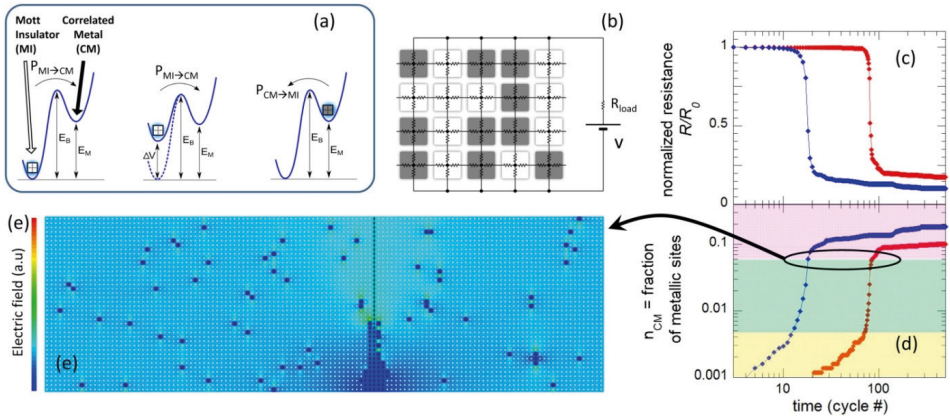


Figure 9: *(a) Energy landscape model used to simulate the Mott IMT driven by an external electric field. This landscape is applied to every cell of the resistor network (b), where grey and white dots represent respectively cells in the MI and CM (transited) states. (c) Resulting simulated evolution of normalized resistance R/R0. The applied voltage is higher for the blue curve than for the red one. (d) Associated increasing fraction of metallic sites in the resistor network. The yellow, green and pink areas correspond respectively to the increase of metallic cells before, during and after the creation of the filamentary percolating path. (e) Representation of the resistor network and associated electric field, just after the creation of this filament. Reproduced with permission from Ref. [123], Copyright 2014, WILEY-VCH Verlag GmbH & Co.*

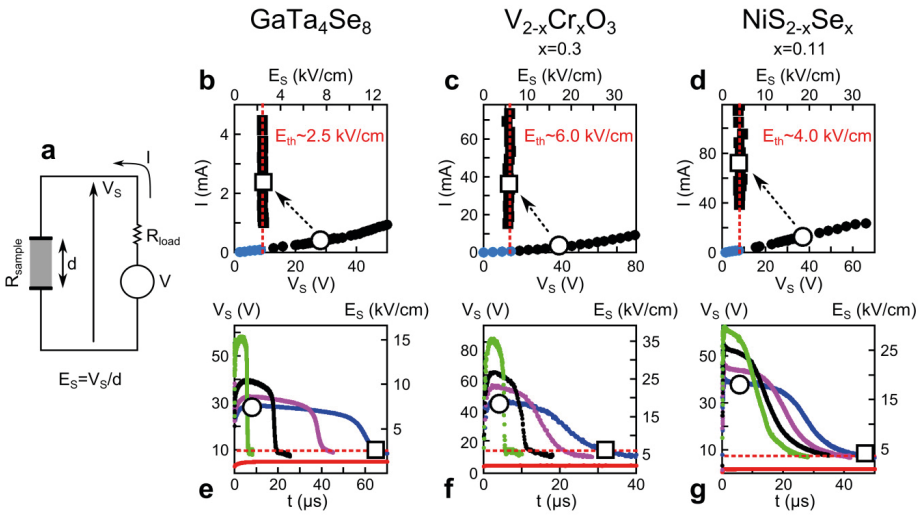


Figure 10. *Panel* ***a*** *shows the schematics of the experimental setup. Universal dielectric breakdown I-V characteristics (top panels* ***b***, ***c***, ***d****) and time dependence of the sample voltage Vs(t) (bottom panels* ***e***, ***f***, ***g****) are displayed for three different types of narrow gap Mott insulators. Blue dots correspond to the region below E_{th}, where no breakdown is observed. Black symbols correspond to the I-V characteristic in the resistive switching region, above E_{th}. The black dots show the initial I-V, before the breakdown, and the black squares indicate the final state. The open symbols highlight a particular breakdown transition for easier visualization. Measurements on $GaTa_4Se_8$ were performed at 77 K [8], on $V_{2-x}Cr_xO_3$ (x=0.3) at 164 K and on $NiS_{2-x}Se_x$ (x=0.11) at 4 K. Reproduced with permission from Ref. [118],Copyright 2013, WILEY-VCH Verlag GmbH & Co.*

4.4 Avalanche phenomena in Mott Insulators: a universal property

The avalanche breakdown phenomenon is a universal property of classical semiconductors. According to the modeling detailed above, avalanche phenomenon should also occur in any Mott insulator provided that the electric field is strong enough to destabilize sufficiently the Mott Insulating state. Recent experiments support that the avalanche breakdown as observed in the AM_4Q_8 compounds can be found in other narrow gap Mott Insulators. Avalanche breakdown was indeed demonstrated in the famous Mott Insulators $(V_{1-x}Cr_x)_2O_3$ and $NiS_{2-x}Se_x$.[118] **Figure 10** shows that these compounds exhibit a similar behavior as $GaTa_4Se_8$ with a sharp transition onset at a threshold electric field of the order of a few kV/cm. In the same way, the avalanche breakdown model might also explain the resistive switchings in Mott insulators like Sr_2CuO_3 and $SrCuO_2$,[93] or κ-(BEDT-TTF)$_2$Cu[N(CN)$_2$]Br[96] as the threshold fields and I(V) characteristics observed for these compounds are quite similar. Avalanche breakdown appears therefore as a universal property of narrow gap Mott insulators. This transition can be considered as a new type of Mott transition. However avalanche breakdown differs from the filling-control or bandwidth-control IMT described in Section 2. These classical Mott transitions are indeed static bulk properties while avalanche breakdown appears as a dynamical and filamentary Mott transition. This electric field controlled IMT will be noted thereafter as type 5.

5 Non Volatile resistive switching in Mott insulators

5.1 Evidence of electric field driven non-volatile Mott IMT

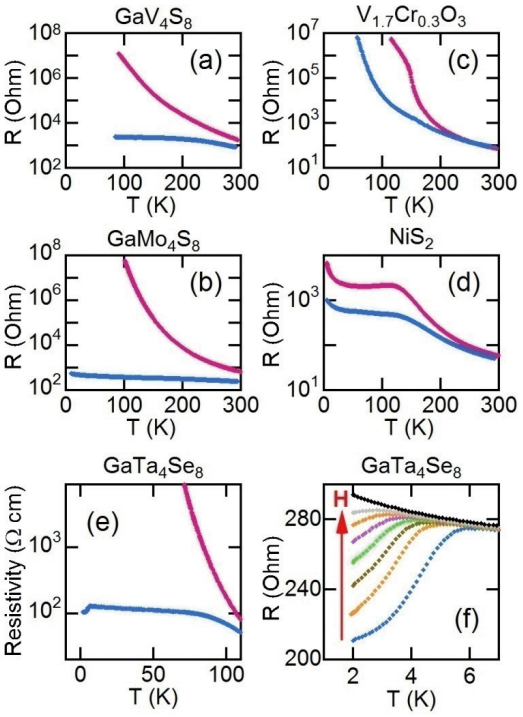


***Figure 11**: Variation of resistance as a function of temperature for various narrow gap Mott insulators in pristine state (pink curves) and after the application of an electric pulse inducing a non-volatile resistive switch (blue curves), in (a) GaV_4S_8, (b) $GaMo_4S_8$, (c) $V_{1.7}Cr_{0.3}O_3$, (d) NiS_2 and (e) $GaTa_4Se_8$. (f) resistance vs. temperature for various magnetic fields (from 0 to 5 Tesla) in a transited $GaTa_4Se_8$ single crystal. [102]*

For electric fields well above the avalanche threshold field involved in the volatile transition, the AM_4Q_8 compounds exhibit a non volatile resistive switching. [13] Indeed, the application on AM_4Q_8 crystals of short voltage pulses of large amplitude induces a non volatile drop of their resistance, namely the low bias resistance measured after the end of the pulse remains at a low resistance value. **Figure 11**a shows the resistance vs. temperature curve of a GaV_4S_8 crystal measured at low bias level in the pristine and transited states. Whereas the pristine curve is typical of an insulator, the transited state is characteristic of a metallic-like material, showing a drop a resistance of several orders of magnitude. As observed for the volatile transition, the non-volatile transition appears in all AM_4Q_8 compounds (see the examples of $GaMo_4S_8$ and $GaTa_4Se_8$ shown in **Figure 11**).[13,120,121] Interestingly, **Figure 11**.c-d demonstrates that this non-volatile resistive switching behavior can also be extended to the other Mott insulators such as $V_{1.7}Cr_{0.3}O_3$ and NiS_2, where the volatile transition has been previously displayed. These recent results suggest that both the volatile and non-volatile resistive switchings could be generalized to the entire class of narrow gap Mott insulators.

Moreover the detailed study of this non-volatile transition has shown that the successive application of unipolar electric pulses to these Mott insulators makes the resistance switch back and forth between high and low resistance states.[13] This reversibility of the non volatile transition enables envisioning memories based on these materials.[122] Noteworthy intermediate levels between high and low resistance states can be reached,[102] which could be of interest for multi-level data storage or memristive applications.

5.2 From volatile to non volatile resistive switching: control of SET and RESET

The existence of a volatile resistive switching (RS) above a threshold electric field which becomes non-volatile at higher field is a specific fingerprint of narrow gap Mott insulators. Recent experiments provide insight into the relationship between these two types of switchings. [123] **Figure 12**a-b show, for example, that a series of seven identical pulses yields a non-volatile transition while each of these pulses applied independently would only trigger a *volatile* resistive switching. Such an evolution from single pulse/volatile RS to multipulse/non-volatile RS can be rationalized on the basis of the model of resistor network with two competing phases already introduced in Section 4.2. This model described on **Figure 9** indeed predicts that the application of an electric field in a Mott insulator induces an accumulation of metallic sites. A *volatile* resistive switching is triggered above a critical accumulation threshold, through the creation of a conductive percolating path. This model also predicts, as shown in the upper part of **Figure 9**d, that the number of metallic sites still goes on increasing after the creation of the filament, as long as the electric field is applied. Simulations show that this accumulation effect corresponds to an increase of the filament diameter. In the experiments described above, the filament diameter is then much larger after application of a series of a few consecutive pulses than after a single pulse. These simulations thus strongly suggest that the observed non-volatile stabilization of the RS ('SET transition') is directly related to the growth of the conducting filament. This concept of critical size above which the filament becomes stable is consistent with classical mechanisms of nucleation and growth processes, where stabilization of a phase becomes possible only above a critical size.[124]

Another appealing prediction of the model of resistor network with competing phases is that the relaxation of metallic domains toward their more stable (Mott) insulating state is thermally activated. This suggests that Joule self-heating could be used to promote the RESET transition to the high resistance state. [123]. **Figure 12**c shows that the application of a very long pulse with electric field chosen to optimize the competition between relaxation (heating effect) and creation (electric field effect) of metallic sites, is indeed efficient to induce the RESET. This long pulse relaxes the resistance to a value very close to the pristine state, which may indicate the quasi-complete dissolution of the filament. On the other hand, no RESET transition is observed when the duration of the pulse is reduced by a factor 4 (see **Figure 12**c) which fully supports a thermal mechanism for the dissolution of the filament. Schemes of the filament evolution suggested by these experiments are shown in **Figure 12**.

To sum up, these experiments demonstrate the relevance of the model of resistor network with two competing phases in the description of both the volatile and non-volatile resistive switchings. According to this model, the volatile resistive switching corresponds to the creation of a conducting filament too thin to be stabilized after the end of the electric field pulse. Conversely, for a non-volatile resistive switching, the thickness of filament is sufficiently large to allow its stabilization after the pulse. **Figure 13** summarizes this scenario and provides schematic representations of the evolution of the filament during the volatile, the "SET" and "RESET" transitions. Finally, a very clear strategy of electric pulses application emerges from this work: applying short multipulses with large electric field for the SET and long single pulses with low electric field to promote the RESET.

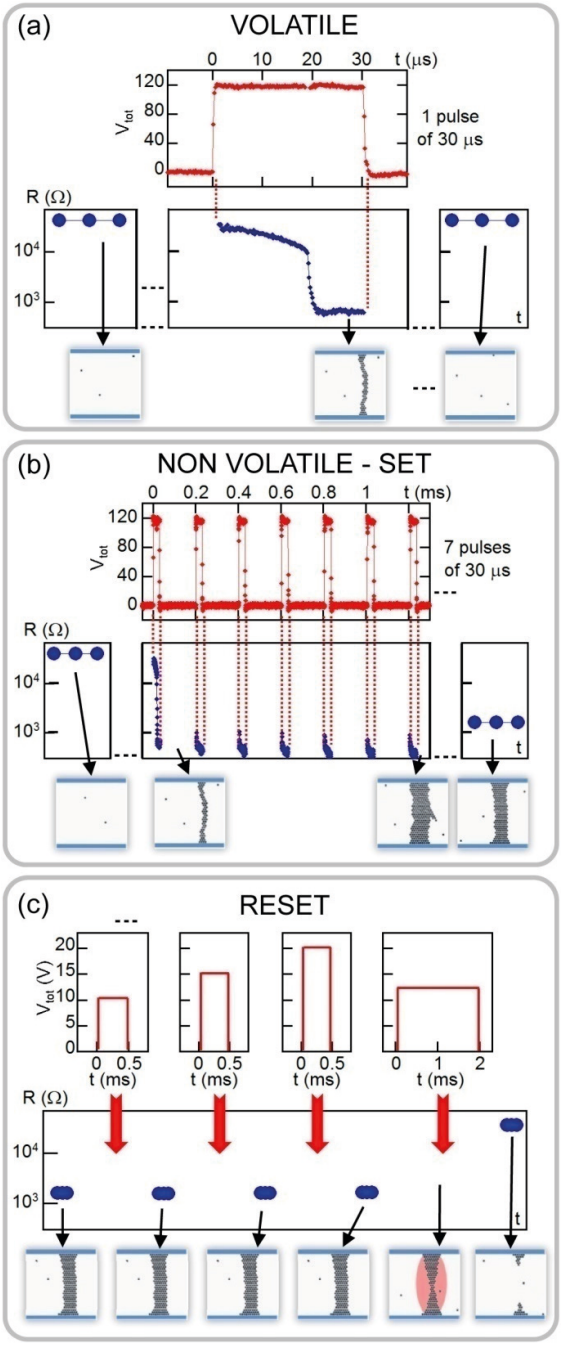


Figure 12 *: Resistance variation of a GaV_4S_8 crystal, before, during and after applying (a) 1 pulse of 30 μs / 120V, and (b) a train of 7 pulses of 30 μs / 120V every 200μs leading to SET non-volatile transition. As expected, the resistance drop during the first pulse shown in (b) is similar to the one occurring during the single pulse in (a). Noteworthy the resistance does not go back to a high resistance state between and during the subsequent pulses. The resistances before and after the application of the (series of) pulses are measured at low bias and are displayed as blue circles. (c) Pulse duration impact on the RESET transition in a GaV_4S_8 crystal. 500 μs pulses in the 10-20 V range do not affect resistance level, whereas a 2 ms / 12V pulse induces the RESET transition. The additional sketches illustrate the evolution of the conductive filament through the application of successive pulses.*

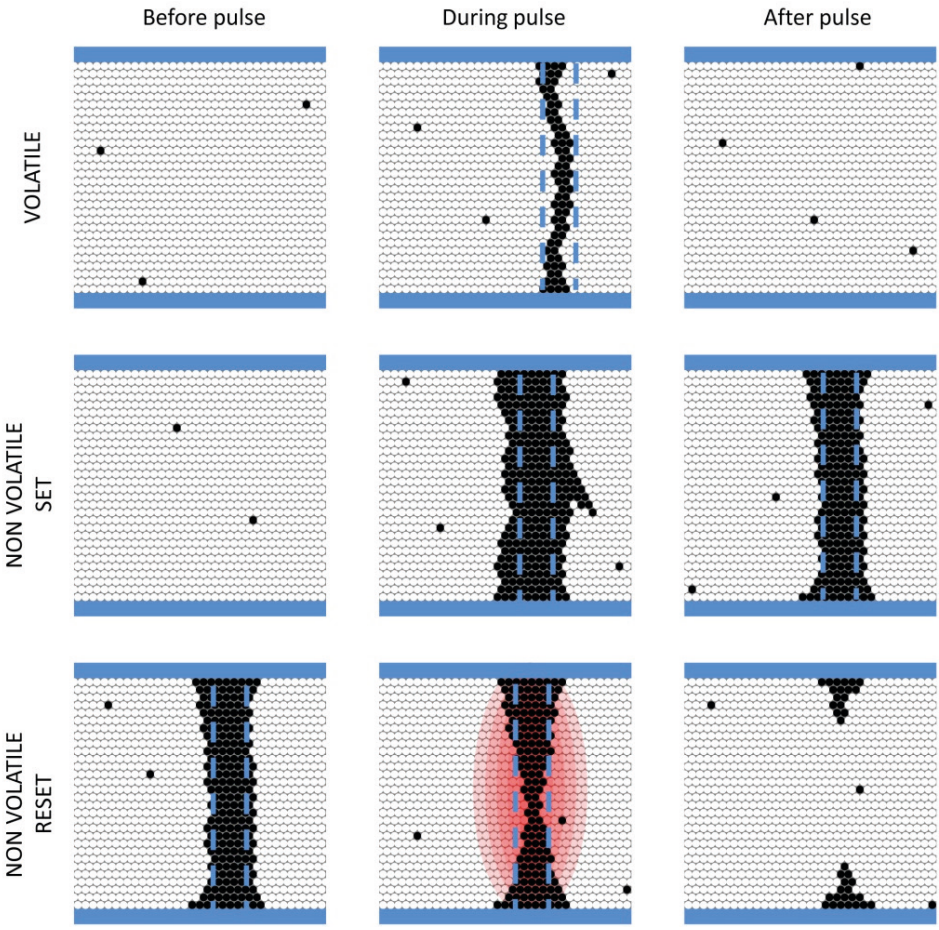


Figure 13 *: Schematic illustration of the filament evolution before, during and after the application of a pulse inducing a volatile transition, a "SET" non-volatile transition and a "RESET" non-volatile transition. White and black domains represent respectively Mott Insulating and Correlated Metal regions of the material. Top and bottom electrodes are depicted in blue, and dashed blue lines represent a critical radius of stability for the filament.*

5.3 Electric-field-induced electronic phase separation and resistive switching

The model of resistor network with competing phases successfully describes key features of a macroscopic property, the volatile resistive switching. It also suggests that new conducting domains should appear at a microscopic level after a non-volatile RS. Scanning Tunnel Microscopy/Spectroscopy (STM/STS) experiments have been carried out on freshly cleaved $GaTa_4Se_8$ single crystals before and after a non-volatile resistive switching to explore this hypothesis. [12,101,125] These experiments have revealed that the non-volatile RS is related to an electronic phase separation at the nanoscale. While the surface topography of pristine crystals is structureless, filamentary structures made of nanoscale heterogeneities with a typical size of 30-70 nm appear after RS, qualitatively oriented along the direction of the electric pulses (**Figure 14**a-b).[12] The analysis of STS map shown in **Figure 14**d reveals that, beside an insulating matrix with STS spectra similar to the pristine state (green areas - curves A in **Figure 14**e-f), these RS-induced nanoscale heterogeneities consist in two different kinds ofdomains. The first ones are metallic (red areas – curve B in **Figure 14**d-f) and the others super-insulating (blue-violet areas, curve C), *i.e.* with a low bias conductance smaller than the pristine one. Moreover the analysis of the tunnel conductance *vs.* voltage measured on each point of **Figure 14**d was used to extract the distribution of the local electronic gaps. Before resistive switching, the distribution of the gap values is homogeneous around 200 meV in the pristine state, as shown in **Figure 14**g. Conversely, **Figure 14**h shows that new gapless regions appear after RS (the metallic regions in red on **Figure 14**d), whereas the super-insulating regions (blue-violet regions in **Figure 14**d) correspond to a continuum of larger gaps between 200 and 700 meV, embedded in an undisturbed matrix whose gap distribution is centered around 200 meV.[125]

The STM experiments demonstrate therefore that for the nonvolatile resistive switching the metallic filamentary paths are made of a percolating granular metallic phase instead of a percolating metallic phase as suggested by the modeling work. This is further confirmed by transport measurements performed after a non volatile resistive switching. The resistance of the crystal is then well described by a two resistance model considering a granular metallic phase (with power law temperature dependence) placed in parallel with an insulating pristine-like phase (with an activated law temperature dependence). [102]

The nanodomains revealed by STM were carefully investigated by Energy Dispersive X-Ray spectroscopy, and by Transmission Electron Microscopy.[120] No chemical composition change nor any crystallographic symmetry breaking or amorphisation between the electrodes were detected at the nanometric scale. This excludes the formation of conducting bridge-like filaments,[126] amorphous-crystalline transition as observed in phase change materials [127] or phase transition similar to the monoclinic-tetragonal phase change observed in VO_2 (see discussion in Section 2.3).

Moreover STM / STS studies have revealed the extreme sensitivity of the crystal surface to the electric field generated by the STM tip. Applying voltages above a threshold value between the STM tip and the surface indeed allows switching nanodomains of typical diameter 10-20 nm. These pristine – metal or pristine – superinsulating switchings are *reversible* and always accompanied by a small topographical change of the surface.[125] For higher tip-surface voltage, the surface deformation is drastically enhanced and leads ultimately to an *irreversible* indentation of $GaTa_4Se_8$ crystal by the STM tip (see Ref. [101] for more details). Both effects are completely unusual and provide clear evidence of a strong electromechanical coupling in $GaTa_4Se_8$.

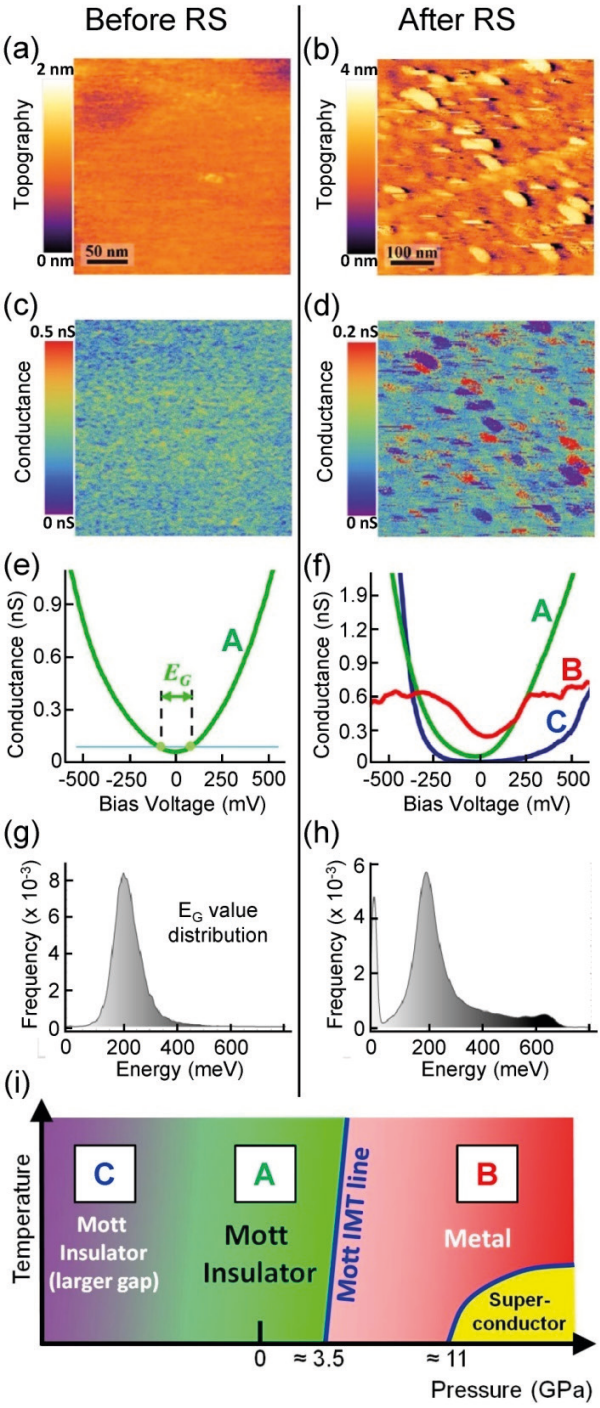


Figure 14 *: STM/STS study of a freshly cleaved $GaTa_4Se_8$ surface, before and after RS : small-scale topographic STM images of (a) the pristine crystal and (b) the transited crystal. (c) Conductance map measured at −200 mV of the area shown in (a) showing a homogeneous electronic state. (d) Conductance map measured at −200 mV of the area shown in (b) exhibiting strong electronic heterogeneities. (e) Representative tunneling conductance spectrum of a pristine cleaved crystal. The gap E_G ∼ 100−200 meV measured by optical and resistivity measurements is indicated by the threshold blue line. (f) Tunneling spectra corresponding to zones A (green), B (blue−violet), and C (red) displayed on image (d). The dI/dV spectra of the zone A (in green) are similar to the one of the insulating pristine samples; the spectra from zone B (blue−violet) are more insulating and hence are called super-insulating while the spectra from zone C (in red) are "metallic-like". (g) and (h) distribution of the electronic gap extracted from a 500×500 nm² STS map, in the pristine (g) and transited state (h). (i) Schematic temperature − pressure phase diagram of the Mott insulator $GaTa_4Se_8$ in its pristine state. For negative pressure (expansion), the Mott−Hubbard gap increases continuously. For positive pressure (compression), a discontinuous first order transition occurs at a critical pressure (≈ 3.5 GPa), and the compound undergoes a Mott IMT. Above ≈11 GPa, $GaTa_4Se_8$ becomes superconducting with critical temperature in the 4-7 K range [46]. Adapted with permission from Ref. [125], 2013, American Chemical Society.*

5.4 Towards a microscopic view of the resistive switching in Mott insulators

The STM/STS experiments have unveiled a particularly important feature of the non-volatile resistive switching in AM_4Q_8, *i.e.* the existence of electric field induced metallic nanodomains without any evidence of Crystallographic Symmetry Breaking (CSB) with respect to the pristine Mott insulating phase. Interestingly, this is reminiscent of the coexistence of phases sharing the same crystallographic structure that develops across the Mott IMT line in Cr-substituted V_2O_3.[128] This absence of CSB in transited AM_4Q_8 thus reminds the behavior expected across the "type 1" Bandwidth Controlled Mott transition discussed in Section 2. This suggests that the metallic domains shown in **Figure 14**d could correspond to *compressed* domains of $GaTa_4Se_8$ which have crossed the IMT line shown in **Figure 14**i.

This hypothesis was tested using the superconducting transition observed at low temperature (T_C= 4-7K) in compressed $GaTa_4Se_8$ above 11 GPa.[46] **Figure 11**e-f show that the resistance drops below 6K in a transited crystal of $GaTa_4Se_8$. This resistance drop is gradually suppressed by a magnetic field of 5T, *i.e.* a value in the same range as the critical field H_{c2} determined on bulk $GaTa_4Se_8$ under pressure.[46] Moreover, the resistance drop displayed in **Figure 11**e-f is only partial and does not go to zero. All these features indicate the presence of granular superconductivity, *i.e.* of disconnected and non-percolating superconducting domains after resistive switching in $GaTa_4Se_8$. The absence of percolation is clearly consistent with the spatial distribution of metallic (red) domains depicted in **Figure 14**d, which are disconnected from each other. To sum up, the presence of granular superconductivity directly proves the presence of compressed metallic domains in transited crystal of $GaTa_4Se_8$.

The existence of compressed (metallic) domain has an interesting consequence: from simple arguments of volume conservation within the $GaTa_4Se_8$ crystal volume, one can infer that expanded domains should coexist with the compressed ones. As discussed in Ref. [25] and shown in **Figure 14**i, expanding a Mott insulator leads to increase its Mott-Hubbard gap. This scenario thus rationalizes the STM/STS studies displayed in **Figure 14**. In particular, the seemingly complex "electronic patchwork" shown in **Figure 14**d simply consists in a set of compressed metallic and neighboring expanded super-insulating domains, embedded in a pristine insulating matrix, and organized along filamentary pathways.

More generally, all these results suggest that the electronic avalanche breakdown induces the collapse of the Mott insulating state into a correlated metallic state. This effect occurs at the local scale and leads to the formation of a granular conductive filaments formed by compressed metallic domains and expanded "superinsulating" domains. This idea that a purely electronic effect, the avalanche, is responsible for a strong response of the lattice is quite natural in the context of Mott IMT physics. For example, the driving force of all Mott IMT is also purely electronic and the lattice response (*e.g.* the volume contraction at the bandwidth controlled IMT)[33] appears a simple consequence of this electronic effect.[26] The Dynamical Mean Field Theory indeed predicts that this lattice response follows from a dramatic change in the electronic wavefunction across the IMT, which has a direct effect on the compressibility of the lattice.[26] The strong sound velocity anomalies reported at the IMT in Cr-substituted-V_2O_3 [129] and in molecular Mott insulators [130,131] provide direct evidence of this effect. The electric-field-induced resistive switching hence appears as a new type of out of equilibrium Mott insulator to metal transition and as a universal property of narrow gap Mott insulators. Finally the modeling work of this original mechanism of RS supplies strategies to control both SET and RESET non volatile transitions. This will be valuable for the realization of efficient ReRAM devices based on narrow gap Mott insulators. The fabrication of such devices and the characterization of their performances are addressed in Section 6.

6 ReRAM devices based on avalanche breakdown in narrow gap Mott insulators

The resistive switching based on electric field controlled IMT discovered on Mott insulator compounds like AM_4Q_8, [12, 13] $V_{1.7}Cr_{0.3}O_3$ or NiS_2 leads to non volatile transitions which makes them potential candidates for ReRAM applications. Studies on this type of ReRAM are scarce and mainly focused on GaV_4S_8. The following sections present therefore the realization of MIM devices using the narrow gap Mott insulator GaV_4S_8 and describe the performances obtained on these devices in the context of ReRAM applications.

6.1 Preparation of GaV_4S_8 thin active layers and GaV_4S_8 based MIM structures

The deposition of GaV_4S_8 material in the form of thin layers has been investigated both by non-reactive RF magnetron sputtering in pure argon [132] and by reactive process in Ar/H_2S mixture [133] using a stoichiometric GaV_4S_8 target [134]. A process parameter window enabling to obtain thin films has been determined both in non-reactive [132] and in reactive gas mixtures [133]. For both approaches, GaV_4S_8 thin films need to be annealed in the 450-600 °C range to exhibit a crystalline structure, as checked by XRD analysis (**Figure 15**a). For films deposited in pure Ar, the annealing is performed with excess sulfur, whereas thin films deposited in reactive phase H_2S/Ar do not need any enrichment to achieve the targeted sulfur stoichiometry. After annealing, the stoichiometric polycrystalline layers crystallize with the expected lacunar spinel structure, (**Figure 15**a), and exhibit a granular morphology as revealed by the SEM image of a 100 nm annealed thick film elaborated with 1% H_2S content (**Figure 15**b). Several Metal Insulator Metal (MIM) structures $Au/GaV_4S_8/Au$ were subsequently realized (**Figure 15**e-f) using these well crystallized GaV_4S_8 thin layers. TEM analyses reveal the excellent crystalline quality of the GaV_4S_8/Au interface at top and bottom electrodes, with GaV_4S_8 atomic planes clearly visible at 2 nm from the interface (**Figure 15**c), without any interfacial amorphous layer [135].

6.2 Resistive switching in GaV_4S_8 MIM structures

Resistive switching experiments were performed on GaV_4S_8 MIM structures. A non-volatile resistive switching can be induced by applying electric pulses to polycrystalline GaV_4S_8 thin films. The resistance *vs.* temperature dependence of the GaV_4S_8 polycrystalline thin layer, displayed in **Figure 16**b, changes from an insulating state in the pristine state (R_{OFF} = red curve) to a conductive one (R_{ON} = blue curve) after the application of short electric pulses in the 500 ns-10 µs range. This is completely similar to the resistive switching observed previously on single crystal (see comparison in **Figure 16** a-b). Moreover, a significant difference between R_{ON} and R_{OFF} is still observable on thin films at 300 K (**Figure 16**b). The pulse protocol described in Section 5.2 was therefore tested at room temperature in order to control the SET and RESET transitions on GaV_4S_8 MIM structures. This voltage pulse protocol alternates a series of seven identical short pulses of large amplitude to generate the SET transition with a single long and low amplitude pulse to generate the RESET transition. Using this pulse protocol a reversible switch back and forth between the high and low resistance states was observed at room temperature on this $Au/GaV_4S_8/Au$ MIM structure (**Figure 16**).

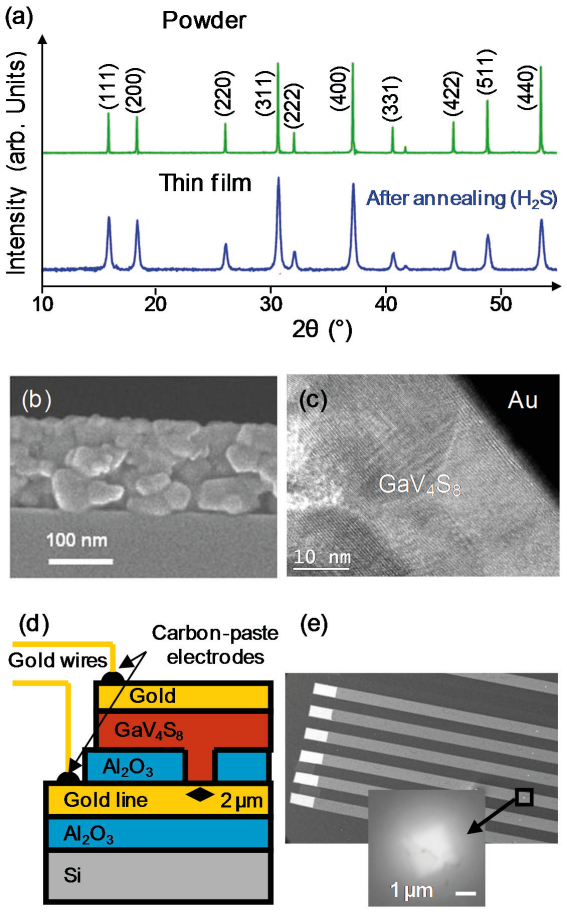


Figure 15 *: Typical characteristics of GaV_4S_8 thin layers (from top to bottom): (a) X-ray diffraction pattern of a 400 nm thick layer after 1 h annealing at 873 K under H2S flow and comparison with the one of home-synthesized powder used as a reference ; (b) SEM image in cross section of a 100 nm thick GaV_4S_8 layer elaborated with 1% H_2S after annealing at 813 K; (c) high magnification TEM picture of the bottom GaV_4S_8/Au interface within a 50×50 µm² Au/GaV_4S_8/Au MIM structure ; (d) Schematic drawing of the 2×2 µm² MIM structure cross-section; (e) SEM images of the corresponding substrate before deposition of the GaV_4S_8 layer (surface view).*

6.3 Performances of GaV_4S_8 based ReRAM devices

Electrical performances such as endurance, scalability, and retention times were evaluated on GaV_4S_8 MIM structures. The endurance was measured on a GaV_4S_8 based device and exceeds 65 000 successive cycles with less than 0.01% error rate [135].

The downscaling properties were also investigated on MIM structures [135] with electrode pad size ranging from 50×50 µm² down to 150×150 nm². As displayed in **Figure 16**d the R_{OFF}/R_{ON} ratio strongly increases with decreasing pad area and reaches values larger than 1000 for pads of 150×150 nm². This result can be easily explained considering the filamentary model depicted in section 5. As long as the cycling involves the creation / full dissolution of a single filament, R_{OFF} is indeed expected to scale with the inverse of the pad area 1/S while R_{ON} is expected to depend only on the resistance of few filamentary conducting paths covering a small area. As a consequence, R_{OFF}/R_{ON} should increase as 1/S for small pads area, which is observed

experimentally for areas below 5 μm² (**Figure 16**d) and should keep increasing as long as pad sizes remain larger than the filamentary conducting paths. R_{OFF}/R_{ON} ratios larger than the 10^3-10^4 current values can thus be expected with further pad size downscaling.

The stability of high and low resistive states obtained on MIM structures has been investigated at room temperature. Extrapolation of R_{OFF} and R_{ON} to 10 years shows respectively slight increase and decrease of these resistance levels (**Figure 16**e-f). Both states exhibit therefore good retention time which is promising for data storage.

Another interesting feature of the switching mechanism observed in narrow gap insulators is that it enables a simple way to tune the SET voltage. Indeed the resistive switching is driven by electric field of the order of kV/cm. The SET voltages used on single crystals (typically 30-50 V for 10-30 μm inter-electrode distance) largely decreases on thin films (down to 1.5 V for 150 nm). SET voltage value lower than 1 V is therefore expected for sub-100 nm thick thin films targeted in future devices. Finally writing time (SET transition) of 7 x 15ns and erasing time (RESET transition) as short as 500 ns were obtained in GaV_4S_8 planar structures. [135]

To summarize, the endurance of Mott-RAM devices is very promising compared to values ranging from 10^3 to 10^7 cycles currently obtained in Flash technology [122]. The writing time of 7x15 ns and the erasing time of 500 ns are favorable compared to characteristics achieved in Flash technology, *i.e.* writing time of 1 μs and even much better than the typical erasing times of 10 ms. In addition, the writing/erasing voltage in the 1 V range stands as a huge advantage when compared to the 12 V reported for Flash memories [122]. Among other ReRAM emerging technologies, Mott insulator based ReRAM devices could be thus considered as really promising candidates to take over the Flash technology.

7 Conclusion

Insulating state may arise in systems with an integer number of unpaired electrons owing to strong electronic correlations. The most prominent examples of these type of systems, known as Mott Insulators, are $(V_{1-x}Cr_x)_2O_3$, $NiS_{2-x}Se_x$ and AM_4Q_8. There are several ways to destabilize the Mott insulating state. The best known ones consist in either applying pressure (Bandwidth-Controlled IMT, type 1), changing the temperature in the vicinity of the Mott transition line (Temperature-controlled IMT, type 2) or doping the system away from half filling (Filling-controlled IMT, type 3). Recently another way to destabilize the Mott insulating state was reported consisting in applying strong electric field. Electric field can indeed initiate a dielectric breakdown of the avalanche type in Mott Insulators which can be considered as an Electric-Field-controlled IMT (type 5). These insulator to metal transitions which emerge from the Mott insulating state are called Mott transitions and do not involve a change in the crystal structure symmetry. Alternatively, many insulating correlated materials like VO_2, Ca_2RuO_4 or Fe_3O_4 display Temperature-controlled IMT (type 4) associated with diverse phase changes that all involve crystallographic symmetry breakings.

Temperature, filling or Electric-Field-controlled IMT which appear in Mott or correlated insulators are interesting in the context of Resistive RAM. Indeed, resistive switchings in Mott or correlated insulators can be classified under three types of mechanisms depending on the type of IMT responsible for the change of resistance (see **Figure 17**). A first type of resistive switching can be explained by a Joule heating induced Temperature-Controlled IMT (type 2 and 4). This thermal mechanism of resistive switching is encountered in Mott insulator systems like $(V_{1-x}Cr_x)_2O_3$ ($x \approx 0.01$) and $NiS_{2-x}Se_x$ ($x \approx 0.45$) and in many correlated insulators like VO_2, Ca_2RuO_4,or Fe_3O_4.

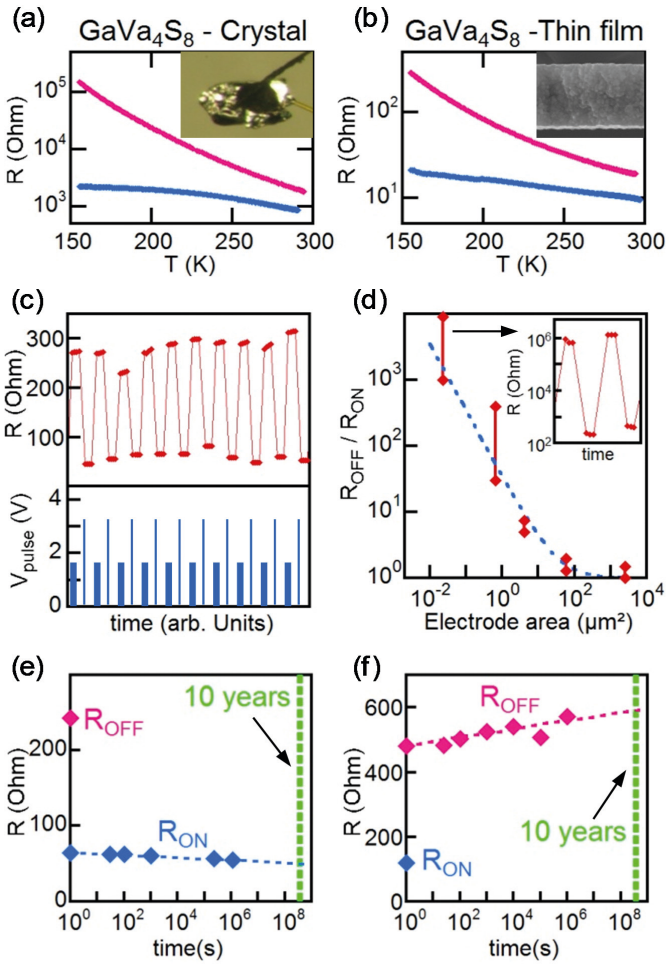


Figure 16 : *Typical electrical characteristics obtained for Au/GaV_4S_8/Au MIM structure exhibiting pad size in the 50 µm to 100 nm range (from top to bottom).*
(a) (b) Comparison of the temperature dependences of high and low resistive states for(a) a 300 µm GaV_4S_8 single crystal and (b) a 400 nm thick GaV_4S_8 thin layer obtained in pure Ar phase (50x50 µm² pad size).
(c) Resistive switching cycles obtained on 2×2 µm² and 150 nm thick GaV_4S_8 based MIM structure in series with R_{load}=10 Ω, by applying successively a multipulse sequence (seven 3.2 V/500 ns pulses, period 3.5 µs) and single 1.6 V/500 µs pulses.
(d) Variation of the R_{OFF}/R_{ON} ratio vs. the electrode area. The dotted line indicates the dependence expected for a simple model of creation / full dissolution of a single filament per memory cell. Inset: RS cycles obtained with a 150x150 nm² pad size exhibiting R_{OFF}/R_{ON} ratio larger than one thousand.
(e) (f) Evolution of R_{ON} (e) and R_{OFF} (f) vs. time for two different 2×2 µm² MIM structures with retention extrapolation to 10 years, and comparison with their initial R_{OFF} and R_{ON} levels.

Insulator to Metal Transition	Resistive switching mechanism		Mott insulators	Correlated insulators
Temperature-Controlled IMT (type 2 and 4)	Thermal		• $(V_{0.99}Cr_{0.01})_2O_3$ [63] • $NiS_{1.45}Se_{0.55}$ [64,65]	• VO_2 [66,67] • NbO_2 [68] • Ca_2RuO_4 [69,5] • V_2O_3 (AFI➔metal) [70,71,72] • Fe_3O_4 [73,74]
Filling-Controlled IMT (type 3)	Valence Change	Filamentary	• NiO [10,86,87] • CuO [8886]	
		Interfacial	• $YBa_2Cu_3O_{7-x}$ [84] • La_2CuO_4 [8]	$Pr_{0.7}Ca_{0.3}MnO_3$ [82]
Electric Field-Controlled IMT (type 5)	Avalanche breakdown induced electronic phase separation		• **$NiS_{2-x}Se_x$ [118]** • **$(V_{1-x}Cr_x)_2O_3$ [118]** • **AM_4Q_8 (A=Ga, Ge; M= V, Nb, Ta, Mo; Q=S, Se, Te)** [12,13,105] • Sr_2CuO_3 [93] • $SrCuO_2$ [93] • κ-$(BEDT\text{-}TTF)_2X$ [96]	

Figure 17 : classification of resistive switching mechanisms in Mott and Correlated Insulators depending on the type of IMT involved in the resistance change. Compounds names written in normal, italic and bold characters display respectively non-volatile, mainly volatile and both volatile/non volatile resistive switching.

A second type of resistive switching observed in correlated and Mott Insulators is based on an ionic migration process. In transition metal oxides migration of oxygen under electric field can indeed induce a filling-controlled IMT (type 3) either along filamentary paths or at the oxide-metal electrode interface. This type of resistive switching first described in band insulators like $SrTiO_3$ is called VCM for Valence Change Memory. Filamentary VCM type resistive switching occurs in Mott insulators like NiO while interfacial VCM type resistive switching occurs for various metal- insulator junctions made of correlated materials like $Pr_{0.7}Ca_{0.3}MnO_3$ or $YBa_2Cu_3O_{7-x}$.

Finally, the last type of resistive switching is related to the Electric-Field-controlled IMT (type 5) or avalanche breakdown recently reported in Mott Insulators. This avalanche breakdown induces the collapse of the Mott insulating state at the local scale and leads to the formation of filamentary conducting paths. Depending on the electric field value these filaments can be either volatile or non-volatile (SET transition). Non-volatile filaments may be destroyed by another electric pulse thanks to Joule heating (RESET transition). This type of resistive switching is universal to narrow gap Mott insulators. It was already demonstrated in several family of Mott insulators like $(V_{1-x}Cr_x)_2O_3$, $NiS_{2-x}Se_x$ and AM_4Q_8. This new mechanism of resistive switching shows promising features such as resistive switching ratio R_{OFF}/R_{ON} exceeding 10^3, cycling endurance reaching more than 65,000 RS cycles, data retention time till 10 years and writing speed below 100 ns. All these results confirm therefore the high potential of this Mott type resistive switching for ReRAM applications.

References

[1] International Technology Roadmap for Semiconductors, Emerging Research Devices (2011) - http://www.itrs.net/

[2] Y. Fujisaki, *Jpn. J. Appl. Phys.* **2013**, *52*, 040001.

[3] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, K. E. Goodson, *Proc. IEEE* **2010**, *98*, 2201.

[4] J.-G. Zhu, *Proc. IEEE* **2008**, *96*, 1786.

[5] F. Pan, S. Gao, C. Chen, C. Song, F. Zeng, *Mater. Sci. Eng. R Rep.* **2014**, *83*, 1.

[6] Yole Développement - i-micronews reports – Emerging non-volatile memory, http://www.i-micronews.com/mems-sensors-report/product/emerging-non-volatile-memory.html

[7] R. Waser and M. Aono, Nature Materials **6**, 833 (2007)

[8] A. Sawa, *Materials today* **2008**, *11*, 28;

[9] R. Waser, R. Dittmann, G. Staikov, K. Szot, *Advanced Materials* **2009**, *21*, 2632.

[10] K. M. Kim, D. S. Jeong, C. S. Hwang, *Nanotechnology* **2011**, *22*, 254002.

[11] D. S. Jeong, R. Thomas, R. S. Katiyar, J. F. Scott, H. Kohlstedt, A. Petraru, C. S. Hwang, *Reports on Progress in Physics* **2012**, *75*, 076502.

[12] C. Vaju, L. Cario, B. Corraze, E. Janod, V. Dubost, T. Cren, D. Roditchev, D. Braithwaite, O. Chauvet, *Advanced Materials* **2008**, *20*, 2760.

[13] L. Cario, C. Vaju, B. Corraze, V. Guiot, E. Janod, *Advanced Materials* **2010**, *22*, 5193.

[14] I. H. Inoue, M. J. Rozenberg, *Adv. Funct. Mater.* **2008**, *18*, 2289.

[15] The site is the entity on which the unpaired electron can be localized. It usually corresponds to a transition metal ion, as *e.g.* in the numerous transition-metal oxides Mott insulators. However the site can also consist in a cluster of atoms, as in AM_4Q_8 chalcogenides compounds, or even in extended molecule as in κ-$(BEDT\text{-}TTF)_2X$.

[16] N. F. Mott, *Proc. Phys. Soc. Sect. A* **1949**, *62*, 416.

[17] J. Hubbard, Proc. R. Soc. Lond. Ser. Math. Phys. Sci. **1964**, 277, 237.

[18] J. Zaanen, G. A. Sawatzky, J. W. Allen, Phys. Rev. Lett. **1985**, 55, 418.

[19] A. Georges, G. Kotliar, W. Krauth, M. J. Rozenberg, Rev. Mod. Phys. **1996**, 68, 13.

[20] G. Kotliar, S. Y. Savrasov, K. Haule, V. S. Oudovenko, O. Parcollet, C. A. Marianetti, *Rev. Mod. Phys.* **2006**, *78*, 865.

[21] G. Kotliar, D. Vollhardt, *Phys. Today* **2004**, *57*, 53.

[22] R. Bulla, T. A. Costi, D. Vollhardt, *Phys. Rev. B* **2001**, *64*, 045103.

[23] The phase diagrams shown in **Figure 1** correspond to the solution of the single band Hubbard Hamiltonian, one of the simplest models of correlated electrons, by the Dynamical Mean-Field Theory (DMFT). The Hubbard model takes into account only the valence electrons moving between lattice sites, with a hopping term *t*

proportional to the bandwidth *W*. Electrons interact between each other only when they are located on the same site, through the local Coulomb repulsion *U*. The DMFT solution becomes exact as the number of neighboring sites increases, and is well suited for two- and three-dimensional experimental systems. See Ref.[19,20,21,22] for more details.

[24] M. Imada, A. Fujimori, Y. Tokura, *Rev. Mod. Phys.* **1998**, *70*, 1039.

[25] Applying a positive external pressure on a Mott insulator induces a volume contraction which enhances the bandwidth *W* and thus reduces the Mott-Hubbard gap $E_G \approx U\text{-}W$. Conversely a volume expansion leads to an enhancement of the gap E_G.

[26] S. R. Hassan, A. Georges, H. R. Krishnamurthy, *Phys. Rev. Lett.* **2005**, *94*, 036402.

[27] A key issue, discussed in Ref. [26], is that electronic degrees of freedom are the (only) driving force of the "type 1" Mott – Bandwidth Controlled Insulator to Metal Transition (IMT). The lattice contraction observed at the type 1 IMT in real systems is therefore only a *consequence* of the softening of electronic degrees of freedom, through the usual electron-phonon coupling.

[28] H. Kuwamoto, J. M. Honig, J. Appel, *Phys. Rev. B* **1980**, *22*, 2626.

[29] A temperature increase corresponds to a vertical path in the phase diagram *T/W vs. U/W* shown in **Figure 1**-a. Indeed, thermal expansion effects in solids usually do not modify significantly neither the bandwidth *W* nor the repulsion *U*, thus keeping the *U/W* ratio constant. Starting from the "PMI" state, there is thus no transition line that can be crossed by increasing temperature.

[30] T. Katsufuji, Y. Taguchi, and Y. Tokura, *Phys. Rev. B* **1997**, *56*, 10145.

[31] P.P. Edwards, T. V. Ramakrishnan, C. N. R. Rao, *The Journal of Physical Chemistry* **1995**, *99*, 5228.

[32] C.H. Yee, L. Balents, *Phys. Rev. X* **2015**, *5*, 021007.

[33] D. B. McWhan, J. P. Remeika, T. M. Rice, W. F. Brinkman, J. P. Maita, A. Menth, *Phys. Rev. Lett.* **1971**, *27*, 941

[34] P. Limelette, A. Georges, D. Jérome, P. Wzietek, P. Metcalf, J. M. Honig, *Science* **2003**, *302*, 89.

[35] D. B. McWhan, T. M. Rice, J. P. Remeika, *Phys. Rev. Lett.* **1969**, *23*, 1384.

[36] A. Jayaraman, D. B. McWhan, J. P. Remeika, P. D. Dernier, *Phys. Rev. B* **1970**, *2*, 3751.

[37] F. Rodolakis, B. Mansart, E. Papalazarou, S. Gorovikov, P. Vilmercati, L. Petaccia, A. Goldoni, J. P. Rueff, S. Lupi, P. Metcalf, M. Marsi, *Phys. Rev. Lett.* **2009**, *102*, 066805.

[38] D. B. McWhan, J. P. Remeika, *Phys. Rev. B* **1970**, *2*, 3734.

[39] P. Limelette, P. Wzietek, S. Florens, A. Georges, T. A. Costi, C. Pasquier, D. Jérome, C. Mézière, P. Batail, *Phys. Rev. Lett.* **2003**, *91*, 016401.

[40] J. M. Honig, J. Spałek, *Curr. Opin. Solid State Mater. Sci.* **2001**, *5*, 269.

[41] H. Benyaich, J. Jegaden, M. Potel, M. Sergent, A. Rastogi, R. Tournier, *J. -Common Met.* **1984**, *102*, 9.

[42] R. Pocha, D. Johrendt, R. Pöttgen, *Chem. Mater.* **2000**, *12*, 2882.

[43] V. Guiot, E. Janod, B. Corraze, L. Cario, *Chem. Mater.* **2011**, *23*, 2611

[44] R. Pocha, D. Johrendt, B. Ni, M. M. Abd-Elmeguid, *J. Am. Chem. Soc.* **2005**, *127*, 8732.

[45] H. Müller, W. Kockelmann, D. Johrendt, *Chem. Mater.* **2006**, *18*, 2174.

[46] M. M. Abd-Elmeguid, B. Ni, D. I. Khomskii, R. Pocha, D. Johrendt, X. Wang, K. Syassen, *Phys. Rev. Lett.* **2004**, *93*, 126403.

[47] R. Pocha, D. Johrendt, B. Ni, M. M. Abd-Elmeguid, *J. Am. Chem. Soc.* **2005**, *127*, 8732.

[48] A. Camjayi, C. Acha, R. Weht, M. G. Rodríguez, B. Corraze, E. Janod, L. Cario, M. J. Rozenberg, *Phys. Rev. Lett.* **2014**, *113*, 086404.

[49] Interestingly, **Figure 3**-b shows that $GaTa_4Se_8$ at 5 GPa (*i.e.* in the metallic phase close to the Mott IMT line) remains metallic down to the lowest temperature. Conversely, pure V_2O_3 displays, in the same region of its phase diagram, a type 4 temperature-controlled metal-insulator transition at ≈ 165 K (see **Figure 2**-d). This difference between $GaTa_4Se_8$ and V_2O_3 demonstrates the non-universal character of the type 4 IMT occurring in pure V_2O_3.

[50] V. Ta Phuoc, C. Vaju, B. Corraze, R. Sopracase, A. Perucchi, C. Marini, P. Postorino, M. Chligui, S. Lupi, E. Janod, L. Cario, *Phys. Rev. Lett.* **2013**, *110*, 037401.

[51] a) E. Dorolti, L. Cario, B. Corraze, E. Janod, C. Vaju, H.-J. Koo, E. Kan, M.-H. Whangbo, J. Am. Chem. Soc. 2010, 132, 5704; b) C. Vaju, J. Martial, E. Janod, B. Corraze, V. Fernandez, L. Cario, Chem. Mater. 2008, 20, 2382. c) B. Corraze, E. Janod, E. Dorolti, V. Guiot, C. Vaju, H.-J. Koo, E. Kan, M.-H. Whangbo and L. Cario, page 116 in Frontiers in Electronic Materials: A Collection of Extended Abstracts of the Nature Conference Frontiers in Electronic Materials, edited by J. Heber and D. Schlom (Wiley, New York (2012).

[52] T. F. Qi, O. B. Korneta, S. Parkin, L. E. De Long, P. Schlottmann, G. Cao, *Phys. Rev. Lett.* **2010**, *105*, 177203.

[53] C. Weber, D. D. O'Regan, N. D. M. Hine, M. C. Payne, G. Kotliar, P. B. Littlewood, *Phys. Rev. Lett.* **2012**, *108*, 256402.

[54] V. Eyert, *EPL Europhys. Lett.* **2002**, *58*, 851.

[55] J. B. Torrance, P. Lacorre, A. I. Nazzal, E. J. Ansaldo, C. Niedermayer, *Phys. Rev. B* **1992**, *45*, 8209.

[56] H. Park, A. J. Millis, C. A. Marianetti, *Phys. Rev. Lett.* **2012**, *109*, 156402.

[57] W.-P. Hsieh, M. Trigo, D. A. Reis, G. A. Artioli, L. Malavasi, W. L. Mao, *Appl. Phys. Lett.* **2014**, *104*, 021917.

[58] H. Wen, L. Guo, E. Barnes, J. H. Lee, D. A. Walko, R. D. Schaller, J. A. Moyer, R. Misra, Y. Li, E. M. Dufresne, D. G. Schlom, V. Gopalan, J. W. Freeland, *Phys. Rev. B* **2013**, *88*, 165424.

[59] M. Dressel, N. Drichko, *Chem. Rev.* **2004**, *104*, 5689.

[60] J. P. Attfield, *Solid State Sci.* **2006**, *8*, 861.

[61] E.J.W. Verwey. *Nature* **1939**, *144*, 327.

[62] J. Li, C. Aron, G. Kotliar, J. E. Han, *ArXiv14100626 Cond-Mat* **2014**.

[63] F. A. Chudnovskii, A. L. Pergament, G. B. Stefanovich, P. A. Metcalf, J. M. Honig, *J. Appl. Phys.* **1998**, *84*, 2643.

[64] F. A. Chudnovskii, A. L. Pergament, P. Somasundaram, J. M. Honig, *Phys. Status Solidi A* **1999**, *172*, 131.

[65] F. A. Chudnovskii, A. L. Pergament, G. B. Stefanovich, P. Somasundaram, J. M. Honig, *Phys. Status Solidi A* **1997**, *161*, 577.

[66] J. Kim, C. Ko, A. Frenzel, S. Ramanathan, J. E. Hoffman, *Appl. Phys. Lett.* **2010**, *96*, 213106.

[67] A. Zimmers, L. Aigouy, M. Mortier, A. Sharoni, S. Wang, K. G. West, J. G. Ramirez, I. K. Schuller, *Phys. Rev. Lett.* **2013**, *110*, 056601.

[68] S. Kim, J. Park, J. Woo, C. Cho, W. Lee, J. Shin, G. Choi, S. Park, D. Lee, B. H. Lee, H. Hwang, *Microelectron. Eng.* **2013**, *107*, 33.

[69] F. Nakamura, M. Sakaki, Y. Yamanaka, S. Tamaru, T. Suzuki, Y. Maeno, *Sci. Rep.* **2013**, *3*.

[70] F. A. Chudnovskii, A. L. Pergament, G. B. Stefanovich, P. A. Metcalf, J. M. Honig, *J. Appl. Phys.* **1998**, *84*, 2643.

[71] J. S. Brockman, L. Gao, B. Hughes, C. T. Rettner, M. G. Samant, K. P. Roche, S. S. P. Parkin, *Nat. Nanotechnol.* **2014**, *9*, 453.

[72] S. Guénon, S. Scharinger, S. Wang, J. G. Ramírez, D. Koelle, R. Kleiner, I. K. Schuller, *EPL Europhys. Lett.* **2013**, *101*, 57003.

[73] T. Burch, P. P. Craig, C. Hedrick, T. A. Kitchens, J. I. Budnick, J. A. Cannon, M. Lipsicas, D. Mattis, *Phys. Rev. Lett.* **1969**, *23*, 1444.

[74] A. A. Fursina, R. G. S. Sofin, I. V. Shvets, D. Natelson, *Phys. Rev. B* **2009**, *79*.

[75] M. D. Pickett, R. StanleyWilliams, *Nanotechnology* **2012**, *23*, 215202.

[76] M.-J. Lee, Y. Park, D.-S. Suh, E.-H. Lee, S. Seo, D.-C. Kim, R. Jung, B.-S. Kang, S.-E. Ahn, C. B. Lee, D. H. Seo, Y.-K. Cha, I.-K. Yoo, J.-S. Kim, B. H. Park, *Adv. Mater.* **2007**, *19*, 3919.

[77] T. Driscoll, H.-T. Kim, B.-G. Chae, M. Di Ventra, D. N. Basov, *Appl. Phys. Lett.* **2009**, *95*, 043503.

[78] S.-H. Bae, S. Lee, H. Koo, L. Lin, B. H. Jo, C. Park, Z. L. Wang, *Adv. Mater.* **2013**, *25*, 5098.

[79] A bipolar resistive switching depends on the polarity of the applied pulse : some filament or interfacial states are created with one polarity and destroyed by the opposite one. Conversely, both polarities have similar effects for a unipolar resistive switching.

[80] D. B. Strukov, G. S. Snider, D. R. Stewart, R. S. Williams, *Nature* **2008**, *453*, 80.

[81] T. Fujii, M. Kawasaki, A. Sawa, Y. Kawazoe, H. Akoh, Y. Tokura, *Physical Review B* **2007**, *75*.

[82] a) S. Q. Liu, N. J. Wu, A. Ignatiev, *Applied Physics Letters* **2000**, *76*, 2749; b) A. Sawa, T. Fujii, M. Kawasaki, Y. Tokura, *Applied Physics Letters* **2004**, *85*, 4073.

[83] H. S. Lee, S. G. Choi, H.-H. Park, M. J. Rozenberg, *Sci. Rep.* **2013**, *3*.

[84] a) M. J. Rozenberg, M. J. Sánchez, R. Weht, C. Acha, F. Gomez-Marlasca, P. Levy, Physical Review B **2010**, 81, 115101; b) C. Acha, M. J. Rozenberg, Journal of Physics: Condensed Matter **2009**, 21, 045702.

[85] the term "Mott insulator" is used here in its broad sense, including both Mott-Hubbard and charge-transfer insulators. See Ref.[18].

[86] D. C. Kim, S. Seo, S. E. Ahn, D.-S. Suh, M. J. Lee, B.-H. Park, I. K. Yoo, I. G. Baek, H.-J. Kim, E. K. Yim, J. E. Lee, S. O. Park, H. S. Kim, U.-I. Chung, J. T. Moon, B. I. Ryu, Applied Physics Letters **2006**, 88, 202102.

[87] K. Kinoshtia, T. Okutani, H. Tanaka, T. Hinoki, K. Yazawa, K. Ohmi, S. Kishida, *Applied Physics Letters* **2010**, *96*, 143505.

[88] Takeshi Yajima, Kohei Fujiwara, Aiko Nakao, Tomohiro Kobayashi, Toshiyuki Tanaka, Kei Sunouchi, Yoshiaki Suzuki, Mai Takeda, Kentaro Kojima, Yoshinobu Nakamura, Kouji Taniguchi, Hidenori Takagi, *Japanese Journal of Applied Physics* **2010**, *49*, 060215.

[89] K. Fujiwara, T. Nemoto, M. J. Rozenberg, Y. Nakamura, H. Takagi, *Jpn. J. Appl. Phys.* **2008**, *47*, 6266.

[90] H. Shima, F. Takano, Y. Tamai, H. Akinaga, I. H. Inoue, *Japanese journal of applied physics* **2007**, *46*, L57.

[91] Lee S B, Chae S C, Chang S H, Liu C, Jung C U, Seo S, Kim D-W, *J. Korean Phys. Soc.* **2007**, *51*, S96.

[92] S. Zhang, S. Long, W. Guan, Q. Liu, Q. Wang, M. Liu, *Journal of Physics D: Applied Physics* **2009**, *42*, 055112.

[93] Y. Taguchi, T. Matsumoto, Y. Tokura, *Physical Review B* **2000**, *62*, 7015.

[94] S. Yamanouchi, Y. Taguchi, Y. Tokura, *Physical Review Letters* **1999**, *83*, 5555.

[95] R. Kumai, Y. Okimoto, Y. Tokura, *Science* **1999**, *284*, 1645.

[96] F. Sabeth, T. Iimori, N. Ohta, *Journal of the American Chemical Society* **2012**, *134*, 6984.

[97] a) T. Oka, R. Arita, and H. Aoki, *Phys. Rev. Lett.* **2003**, *91*, 066406; b) T. Oka and H. Aoki, *Phys. Rev. B* **2010**, *81*, 033103; c) T. Oka and H. Aoki, *Phys. Rev. Lett.* **2005**, *95*, 137601.

[98] F. Heidrich-Meisner, I. González, K. A. Al-Hassanieh, A. E. Feiguin, M. J. Rozenberg, and E. Dagotto, *Phys. Rev. B* **2010**, *82*, 205110;

[99] M. Eckstein, T. Oka, P. Werner, *Physical Review Letters* **2010**, *105*, 146404.

[100] H. Aoki, N. Tsuji, M. Eckstein, M. Kollar, T. Oka, P. Werner, *Rev. Mod. Phys.* **2014**, *86*, 779.

[101] V. Dubost, T. Cren, C. Vaju, L. Cario, B. Corraze, E. Janod, F. Debontridder, D. Roditchev, *Adv. Funct. Mater.* **2009**, *19*, 2800.

[102] C. Vaju, L. Cario, B. Corraze, E. Janod, V. Dubost, T. Cren, D. Roditchev, D. Braithwaite, O. Chauvet, Microelectronic Engineering **85**, 2430 (2008).

[103] The circuit used for resistive switching experiments is schematized in Figure 8a.

[104] M. E. Levinshteĭn, J. Kostamovaara, S. Vainshtein, *Breakdown Phenomena in Semiconductors and Semiconductor Devices*; World Scientific, 2005.

[105] V. Guiot, L. Cario, E. Janod, B. Corraze, V. Ta Phuoc, M. Rozenberg, P. Stoliar, T. Cren, D. Roditchev, *Nat Commun* **2013**, *4*, 1722.

[106] J. L. Hudgins, G. S. Simin, E. Santi, M. A. Khan, *IEEE Trans. Power Electron.* **2003**, *18*, 907.

[107] J. L. Hudgins, *J. Electron. Mater.* **2003**, *32*, 471.

[108] F. Klappenberger, K. F. Renk, R. Summer, L. Keldysh, B. Rieder, W. Wegscheider, Appl. Phys. Lett. **2003**, 83, 704.

[109] H. Fröhlich, On the theory of dielectric breakdown in solids. *Proc. R. Soc. A* **1947**, *188*, 521.

[110] S. Whitehead, *Dielectric breakdown of solids.* (Clarendon Press, 1951).

[111] H. Fröhlich, Theory of electrical breakdown in ionic crystals. *Proc. R. Soc. A* **1937**, *160*, 230.

[112] F. Seitz, On the theory of electron multiplication in crystals *Phys Rev B*. **1949**, *76*, 1376.

[113] H. Fröhlich, F. Seitz, Notes on the theory of dielectric breakdown in ionic crystals. *Phys. Rev*. **1950**, *79*.

[114] H. Eskes, M. B. J. Meinders, and G. A. Sawatzky, *Phys. Rev. Lett.* **1991**, *67*, 1035.

[115] M. B. J. Meinders, H. Eskes, and G. A. Sawatzky, *Phys. Rev. B* **1993**, *48*, 3916.

[116] Photoinduced phase transition, World Scientific publishing, Ed. K. Nasu (2004).

[117] C. Ye, P. Cai, R. Yu, X. Zhou, W. Ruan, Q. Liu, C. Jin, Y. Wang, *Nat. Commun.* **2013**, *4*, 1365.

[118] P. Stoliar, L. Cario, E. Janod, B. Corraze, C. Guillot-Deudon, S. Salmon-Bourmand, V. Guiot, J. Tranchant, M. Rozenberg, Advanced Materials **2013**, 25, 3222.

[119] P. Stoliar, M. Rozenberg, E. Janod, B. Corraze, J. Tranchant, L. Cario, *Physical Review B* **2014**, *90*, 045146.

[120] B. Corraze, E. Janod, L. Cario, P. Moreau, L. Lajaunie, P. Stoliar, V. Guiot, V. Dubost, J. Tranchant, S. Salmon, M.-P. Besland, V. T. Phuoc, T. Cren, D. Roditchev, N. Stéphant, D. Troadec, M. Rozenberg, *Eur. Phys. J. Spec. Top.* **2013**, *222*, 1046.

[121] M. Querré, B. Corraze, E. Janod, M. P. Besland, J. Tranchant, M. Potel, S. Cordier, V. Bouquet, M. Guilloux-Viry, L. Cario, *Key Eng. Mater.* **2014**, *617*, 135.

[122] *International Technology Roadmap for Semiconductors* **2013**. www.itrs.net, in Emerging Research Devices and in Emerging Research Materials.

[123] J. Tranchant, E. Janod, B. Corraze, P. Stoliar, M. Rozenberg, M.-P. Besland and L. Cario, in *Phys. Status Solidi A* **2014**, *212*, 239.

[124] A. Umantsev, *Phys. Nonlinear Phenom.* **2007**, *235*, 1.

[125] V. Dubost, T. Cren, C. Vaju, L. Cario, B. Corraze, E. Janod, F. Debontridder, D. Roditchev, *Nano Lett.* **2013**,*13*, 3648.

[126] C. Schindler, S. C. P. Thermadam, R. Waser, M. N. Kozicki, *IEEE Trans. Electron Devices* **2007**, *54*, 2762.

[127] A. Pirovano, A. Lacaita, A. Benvenuti, F. Pellizzer, R. Bez, *IEEE Trans. Electron Devices* **2004**, *51*, 452.

[128] S. Lupi, L. Baldassarre, B. Mansart, A. Perucchi, A. Barinov, P. Dudin, E. Papalazarou, F. Rodolakis, J.-P. Rueff, J.-P. Itié, S. Ravy, D. Nicoletti, P. Postorino, P. Hansmann, N. Parragh, A. Toschi, T. Saha-Dasgupta, O. K. Andersen, G. Sangiovanni, K. Held, M. Marsi, *Nat. Commun.* **2010**, *1*, 105.

[129] S. Populoh, P. Wzietek, R. Gohier, P. Metcalf, *Phys. Rev. B* **2011**, *84*, 075158.

[130] D. Fournier, M. Poirier, M. Castonguay, K. D. Truong, *Phys. Rev. Lett.* **2003**, *90*, 127002.

[131] M. de Souza, A. Brühl, C. Strack, B. Wolf, D. Schweitzer, M. Lang, *Phys. Rev. Lett.* **2007**, *99*, 037003.

[132] E. Souchier, M.-P. Besland, J. Tranchant, B. Corraze, P. Moreau, R. Retoux, C. Estournès, P. Mazoyer, L. Cario, E. Janod, *Thin Solid Films*, **2013**, *533*, 54.

[133] J. Tranchant, A. Pellaroque, E. Janod, B. Angleraud, B. Corraze, L. Cario, M.-P. Besland *J. Phys. D: Appl. Phys.* **2014**, *47*, 065309.

[134] E. Souchier, L. Cario, B. Corraze, P. Moreau, P. Mazoyer, C. Estounes, R. Retoux, E. Janod, M.P. Besland, *Phys. Status Solidi RRL*, **2011**, *5*, 53.

[135] J. Tranchant, E. Janod, L. Cario, B. Corraze, E. Souchier, J.-L. Leclercq, P. Cremillieu, P. Moreau, M.-P. Besland, *Thin Solid Films* **2013**, *533*, 61.

D 6 Phase Change Materials

Matthias Wuttig
I. Institute of Physics (IA), RWTH Aachen and
Forschungszentrum Jülich GmbH

Contents

1 Introduction and Basics **2**

2 Structure and Bonding **5**
2.1 Atomic Structure of the Crystalline Phase 6
2.2 Electronic Structure of the Crystalline Phase 10
2.3 Atomic Structure of the Melt 15
2.4 Atomic Structure of the Amorphous Phase 16

3 Material Properties **17**
3.1 Vibrational and Thermal Properties 17
3.2 Electrical Properties 19

4 Applications and Outlook **21**
4.1 Optical Storage 21
4.2 Electronic Storage 21
4.3 Other Applications 23

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction and Basics

Phase-Change Materials can rapidly and reversibly be switched between an amorphous and a crystalline phase. Since both phases are characterized by very different optical and electrical properties, these materials can be employed for rewritable optical and electrical data storage. Hence, there are considerable efforts to identify suitable materials, and to optimize them with respect to specific applications. Design rules which can explain why the materials identified so far enable phase-change based devices would hence be very beneficial.

The present chapter describes materials that have been successfully employed and discusses common features regarding both typical structures and bonding mechanisms. It is shown that typical structural motifs and electronic properties can be found in the crystalline state that are indicative for resonant bonding, from which the employed contrast originates. The occurrence of resonance is linked to the composition, thus providing a design rule for phase-change materials. This understanding helps to unravel characteristic properties such as electrical and thermal conductivity which are discussed in the subsequent section. Finally, present approaches for improved high-capacity optical discs and fast non-volatile electrical memories that hold the potential to succeed present-day‘s Flash memory, are presented.

Throughout the history of mankind, the ability to store, share, develop, rearrange and combine information has been key to its evolution. We may only guess what course history would have taken if apparently simple inventions such as pen, paper and letterpress, but also the modern wonder of computers would not be omnipresent. A particularly promising approach for such data storage devices is based on the fast reversible switching of so-called phase-change materials between an amorphous and a crystalline state. Both phases are characterized by very different material properties, thus providing the contrast required to distinguish between logical states. Phase-change recording was initiated in the 1960s by Ovshinsky [3]. It is the state-of-the-art technique for rewritable optical storage since the market introduction of the CD-RW in 1996, and continues to hold this position up to now in the form of the rewritable incarnation of the Blu-ray disc format (i.e., BD-RE). It is also among the most promising candidates to succeed Flash memory, with the potential to fulfill its duties at a speed currently reached by the volatile DRAM. However, phase-change memories are far from being just an engineering issue. A wide scope of aspects in material science is involved in understanding phase-change materials, founding the scientific interest in these materials that has led to hundreds of publications. Ultimately, an in-depth understanding of the few known successful phase-change materials is commonly thought to lead to design rules for optimal materials, which is one of the driving forces of the field. It is the aim of this chapter to provide an overview over the current state of understanding of phase-change materials that has greatly improved within the last few years.

To start, a brief introduction into the basics of phase-change recording shall be given. The principle of operation is illustrated in Figure 1. A small portion of a phase-change material is switched between the crystalline and amorphous state by providing a precisely controlled amount of heat. Currently, either laser pulses or electrical pulses are employed as heat sources depending on the application in optical or electronic data storage. Starting from a crystalline bit, the temperature needs to be first elevated above the liquidus temperature T_l using a short, high intensity (high current) pulse. Since only a spatially confined region is heated up, a huge temperature gradient between the molten bit and the surrounding material is obtained, leading to high cooling rates of about 10^{10} K/s. If the melt cools fast enough, crystallization is bypassed.

Instead a melt-quenched amorphous bit is formed if the temperature falls fast enough below a critical temperature, the glass transition temperature T_g. In this low temperature regime, the atomic mobility is so small that crystallization, though energetically favorable, is kinetically hindered. To switch from the amorphous back to the crystalline state, the temperature of the bit needs to be elevated for a sufficiently long time to a temperature where the atomic mobilities are high enough for crystallization to occur. Hence, the sample has to be heated significantly above the glass transition temperature. To read out whether a bit is amorphous or crystalline, low intensity (low current) pulses are employed to distinguish between low and high reflectivity (conductivity).

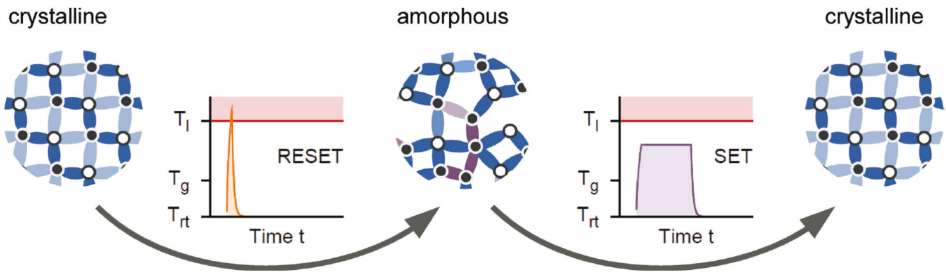


Fig. 1: *The operation principle of phase-change devices is based on the reversible switching between the crystalline and amorphous state. Amorphization (also called RESET -operation) of a bit proceeds via melt-quenching, employing short current pulses as heat sources. In optical recording, short laser pulses are utilized instead. The resulting huge temperature difference between the confined melt and the surrounding material leads to extremely high cooling rates. Thus, the disorder of the liquid is frozen in. Crystallization (SET -operation) requires annealing of an amorphous bit at a temperature below the melting temperature for the atoms to adopt the energetically favorable crystalline order. Reproduced from Ref. [2].*

Noteworthy are the timescales of phase changes. Typically, crystallization is the slowest process involved. Nevertheless, under optimal conditions it may proceed in a matter of nanoseconds at elevated temperatures. At ambient conditions, however, crystallization of an amorphous bit must not take place within many years to ensure data retention. This means that the crystallization rate of phase-change materials must increase by up to twenty orders of magnitude while the temperature is increased by only a few hundred Kelvin. Thus, an in-depth understanding of the crystallization kinetics is required to find out where this significant temperature-dependence stems from and how it can be controlled.

The task to be accomplished for material researchers is to find out which materials are suitable for application. This is a challenging quest, since the requirements regarding kinetics are not the only ones phase-change materials have to meet. These requirements for storage applications are listed in Table 1.

In order to distinguish between the phases, reflectivity and conductivity must differ significantly. At first sight, this seems to require also a significant difference in local structure. But then the question arises, why the phase transition is so fast. The understanding of this apparent conflict has made significant progress in the recent past and is addressed in the following sections. Phase-change materials must provide a compromise of numerous, sometimes conflicting

demands. In the past, a number of materials have been identified, which meet these requirements. Since these materials have often been found by an empirical approach, it is not clear whether these materials are really best suited to meet the demands listed in Table 1.

Application Requirement	Material Requirement
Distinguishable logical states	ability to switch between phases that exhibit significantly different optical/electrical properties
Fast data transfer rates	rapid crystallization at elevated temperatures
Stability	no crystallization at ambient temperatures
Scalability/Data density	simple compositions, insensitivity to composition variations, functionality retained in small volumes with large interfaces
Cyclability	no irreversible modifications due to the switching process itself (e.g., electromigration), intrinsic change of the material over time (e.g., relaxation) or interaction with the environment (e.g., chemical reaction with the electrodes)
Operation of PCRAM at low voltages	nonlinear electrical behavior in the amorphous state and reasonably high resistivity in the crystalline state
Easy to fabricate	production compatible with established semiconductor manufacturing processes and compliant with environmental legislation

Table 1: *Numerous material requirements for phase-change materials follow from the desired use in data storage applications. This table compiles the most important demands. The task material research is confronted with is to find optimal materials that suit these needs.*

Most of the families of phase-change materials already identified can be found in the ternary Ge:Sb:Te-phase diagram shown in Figure 2. The most prominent materials such as $Ge_2Sb_2Te_5$ are located on the pseudo-binary line connecting GeTe and Sb_2Te_3. Besides Sb_2Te_3, also Sb_2Te offers suitable properties when combined with silver and indium, yielding the widely employed AgInSbTe (abbrev. AIST). Finally, another material family that has attracted considerable interest in the last years is found here, namely modifications of antimony such as $Ge_{15}Sb_{85}$. It stands out due to the fact that it does not contain a chalcogenide component, but can be understood in terms of doped antimony. Doping in the field of phase-change materials refers to much larger concentrations (typically on the order of some percent) than in usual semiconductors such as silicon. We will tackle the role of stoichiometry in more detail later on.

The discussions presented above immediately raise the question why the materials in the ternary Ge:Sb:Te-phase diagram work as phase-change materials, and whether materials, that are composed of other elements, would work as well, or even better. While this question can in part be tackled by empirical search algorithms, a microscopic understanding could provide a superior route to material optimization. This might help to understand how both the switching speed and the property contrast can be optimized simultaneously by stoichiometry variation. How can this goal be reached?

If you want to understand function, you have to study structure. This famous expression by Sir Francis Crick (thanks to Professor R. Jones for bringing this quotation to our attention) serves as the guideline for the first half of this article. Presumably, the significant difference in optical

reflectivity and electrical conductivity between the amorphous and the crystalline state is related to substantial differences in the atomic arrangement. Hence, we will first review the structures of crystalline phase-change materials, and subsequently relate them to the bonding. The comparison to the structure and bonding in the amorphous phase will then allow us to identify the origin of the contrast that these materials exhibit. Understanding structure and bonding then provides an efficient framework to discuss other physical properties that are relevant for the materials and their prospective applications. Finally, we conclude with an overview of present and future applications of phase-change materials.

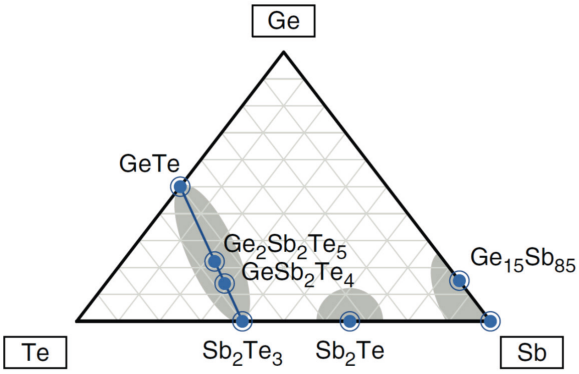


Fig. 2: *Most phase-change materials are found within the ternary Ge:Sb:Te-phase diagram. In particular, the pseudo-binary line between GeTe and Sb_2Te_3 stands out as it hosts the most prominent phase-change materials composed as $(GeTe)_m(Sb_2Te_3)_n$, with m and n being integer numbers. Reproduced from ref. [2].*

2 Structure and Bonding

This chapter will start with a discussion of the atomic arrangement in crystalline phase-change materials. It will be demonstrated that these materials are characterized by a few typical structural motifs, which can be linked to a unique bonding mechanism which prevails in the crystalline phase. Hence, the precise characterization of the atomic arrangement of the crystalline state can provide valuable insight into the bonding of these materials and the resulting properties.

The determination of the structure should be much easier for the crystalline phase than for both the liquid and the amorphous state. The long range order of the crystalline phase enables us to employ the tools of crystallography, with which the precise determination of atomic arrangements is routinely feasible even for crystalline materials having large unit cells and containing several different atomic species. Hence, it appears surprising that even today there are important new discoveries as well as open questions regarding the atomic arrangement of crystalline phase-change materials. Therefore, we will first briefly summarize the challenges to determine the structure of crystalline phase change materials. As already mentioned in the introduction, one of the attractive properties of phase change materials is their ability to crystallize on time scales of just a few nanoseconds at elevated temperatures. The underlying reasons for the fast crystallization behavior will be described in the chapter by M. Salinga (D 7). The specific crystallization characteristics of phase-change materials lead to the fact that their crystalline state usually consists of rather small (few tens of nanometers) grains. Usually, these grains form a polycrystalline state without any preferred orientation. This is a serious complication for diffraction tools, which are ideally suited for single crystals.

Therefore in recent years it has been attempted to produce single crystalline samples of phase change materials in the Ge:Sb:Te-system. A second promising approach to obtain samples with single crystalline qualities has been pursued as well; epitaxial thin films of $Ge_2Sb_2Te_5$ grown on single crystal substrates to overcome the limitations that arise from polycrystalline samples, such as the inability to angularly resolve the Brillouin-zone.

2.1 Atomic Structure of the Crystalline Phase

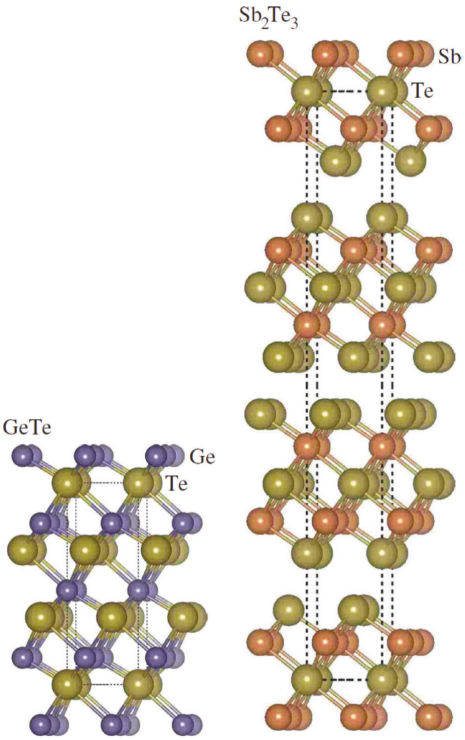


Fig. 3: *The crystal structures of GeTe and Sb_2Te_3 as explained in the text. Phase-change materials inherit the principal structural features of these two limiting cases.*

We now turn to the discussion of typical crystal structures of phase-change materials, focusing again on the prototype Ge:Sb:Te-materials. Interesting enough, almost independent of the formation routes chosen, the atomic structures of phase-change materials exhibit generic features and structural motifs, which are quite different from other materials that at first sight should behave similarly. This implies that phase-change materials are characterized by a unique bonding mechanism, which will be discussed in the following section. The two limiting cases of the pseudo-binary line, GeTe and Sb_2Te_3, are well suited to discuss the structure and bonding mechanisms that rule phase-change materials, see also Figure 3.

2.1.1 The Structure of GeTe

Let us first consider GeTe. It exhibits a structure that closely resembles the rocksalt-structure, with Ge occupying the cation and Te the anion sublattice. At temperatures below approximately 700K, atomic displacements deform this lattice. First, there is a relative shift of the two sublattices along the [111]-direction. This reduces the number of bonds from six to three short bonds (and three long bonds), in line with the Peierls-model of distortions for a system with an average

number of three p-electrons. As a secondary effect, the atomic displacements lead to a small decrease of the cell angle. Altogether, a rhombohedral rather than a cubic cell is obtained.

2.1.2 The Role of Peierls-like Distortions

Since the distortions mentioned above are of utmost importance for phase-change materials, we shall briefly introduce the underlying concept. Though we adopt the common approach to introduce the systematics for a periodic (i.e., crystalline) system, the results on the shifts in the density of states and the local distortions can be transferred to non-crystalline systems as well. As a prerequisite, dominant p-electron bonding (equivalent to the absence of a pronounced hybridization between s- and p- states) is assumed, leaving us first with a (hypothetic) highly symmetric six-fold coordination. To simplify matters, we assume a one-dimensional chain of equally spaced atoms with a simple parabolic band structure as shown in Figure 4. Depending on the number of electrons or equivalently, the position of the Fermi wavevector k_F, this chain is unstable against a periodic distortion. Given k_F is located within the Brillouin zone, that is

$$k_F = \gamma \cdot \frac{G}{2}, \gamma < 1 \tag{1}$$

a distortion that introduces a Fourier component of the potential at this wavevector, V_{kF}, leads to the opening of a gap of size $2|V_{kF}|$ and thus a decrease of the energy of the occupied states. The Brillouin zone shrinks according to the periodicity of the distortion pattern. With this model, we can understand the coordination of GeTe and elemental antimony, for instance, if we assume that each dimension may be treated separately. Then, we have three p-valence electrons per atom out of a maximum of six, one per dimension, which yields $\gamma = 1/2$. Thus, the initially equally long bonds in one dimension split into a short and a long bond. So in three dimensions, three short and three long bonds per atom result. The same line of reasoning can be employed for elemental tellurium, yielding two short and four long bonds per atom and reproducing the experimentally obtained zig-zag chain-structure. However, there is one degree of freedom per dimension, the phase of the distortion that cannot be inferred from the preceding arguments. We note that so far we have not considered additional elastic forces nor the effect of ionicity, that comes into play for non-elemental systems, both counteracting the mechanism

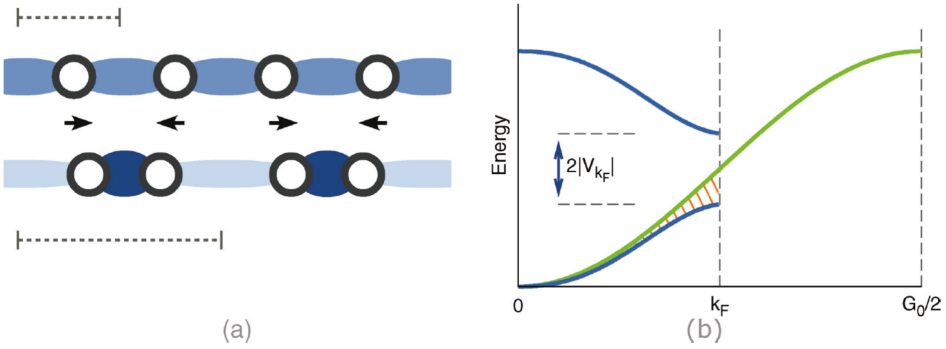


Fig. 4: *Schematic of a Peierls-distortion for a one-dimensional chain with a half-filled band. By a (periodic) distortion, the size of the Brillouin zone is reduced and a new component of the potential is introduced. This leads to an opening of a gap and concurrently a decrease in energy for the occupied states that is the driving force behind these distortions. Reproduced from ref. [2].*

of atomic distortions. As mentioned before, Peierls-like distortions can also prevail in non-crystalline systems. The conceptual translation from a periodic to a local picture can be given by local hybridization of the atoms, thus maximizing the degree of saturation of a few short covalent bonds at the expense of less saturated long bonds.

2.1.3 The 8 - N-rule

The above coordination numbers of antimony and tellurium agree with what would have been predicted by the so-called 8 - *N*-rule. With *N* being the number of valence electrons of an atom, the rule states that generally, the coordination of an atom in a covalent bonding configuration is equal to 8 - *N*. The 8 - *N*-rule successfully explains and predicts the atomic coordination in a wide range of materials composed of elements from the groups V, VI and VII. However, we note that as one goes down in germanium's row of the periodic table, the limits of the 8 - *N*-rule become obvious. Elemental lead is not tetrahedrally, but octahedrally coordinated, owing to the fact that due to relativistic effects hybridization becomes unfavorable. This 8 - *N*-rule has to be applied either individually for each atomic species or in a species-averaged manner. An example for the latter behavior is seen in GaAs. On average, this material has 4 valence electrons (As has 5 valence electrons, while Ga has 3). In this case, it can form a tetrahedral atomic arrangement as in the zincblende structure, where every atom has four nearest neighbors and forms a tetrahedral atomic arrangement. In SiO_2, on the contrary, the 8-N rule would need to be applied for each atomic species individually; silicon, with its 4 valence electrons, has 4 nearest neighbors of oxygen, while oxygen, with its 6 valence electrons, has 2 nearest neighbors of Si. For GeTe, we see that we need to use an *average* number of five valence electrons per atom. The structures we observe here (and in the following), are remarkably close to six-fold coordination, and the validity of the 8 - *N*-rule is only just established by the slight local atomic distortions. This raises the question whether it is reasonable to argue that the 8 - *N*-rule is fulfilled, if the difference between nearest- and next-nearest neighbors becomes very small.

2.1.4 The Structure of Sb_2Te_3

Not only the structure of GeTe, but also that of Sb_2Te_3 can be understood in terms of a distorted rocksalt-like structure. Again, there is an atomic alternation as antimony has only tellurium neighbors in an octahedral environment. However, since there is an excess of Te-atoms, these would need to be arranged in neighboring sites. Instead, well separated layers, consisting of a sequence Te-Sb-Te-Sb-Te, form with interlayer bonding between adjacent Te planes being ascribed to Van der Waals-interaction. Their separation can be explained by electrostatic repulsion of the anionic Te-atoms. For the following discussion, it will be instructive to view the space between two layers as occupied by a layer of intrinsic vacancies. The layer periodicity depends on the Sb to Te-ratio as investigated by x-ray diffraction.

2.1.5 The Structure of alloys on the pseudobinary GeTe-Sb_2Te_3 line

The Ge:Sb:Te-compounds inherit the aforementioned structural ingredients. In the metastable crystalline phase, they exhibit a rocksalt-like structure, with tellurium occupying the anion sublattice. The other, cation sublattice is occupied by germanium, antimony and intrinsic vacancies. Thus, these metastable phases also feature an octahedral-like coordination. The atomic positions show (Peierls-like) distortions from the high-symmetry-positions. For compositions close to GeTe, also a rhombohedral distortion of the unit cell is observed. Interesting enough, the occupation of the cation sublattice depends on the thermal history. After crystallization, typically a random, chemically disordered occupation is obtained. Density functional theory calculations have been performed in order to find out whether there is an energetically preferred occupation. The principal finding is that certain layer-sequences are favored, with the layers being orientated normal to the [111]-direction. The layer sequence is typically similar to what

is found for Sb_2Te_3. The germanium-incorporation leads in most cases to the formation of additional -Te-Ge-Te- sequences within the layer. The initially randomly dispersed vacancies are shifted to form the vacant space in between two layer-stacks.

2.1.6 The Role of Intrinsic Vacancies

Considerable attention has been devoted to the investigation of the large concentration of intrinsic vacancies. The formation of a vacancy in a solid corresponds to the creation of a defect which usually requires a large formation energy of several eV. This explains why most semiconductors such as Si, Ge, or GaAs are characterized by very small vacancy concentrations. On the contrary, in a phase change material such as $GeSb_2Te_4$, there is a 25% concentration of vacancies on the sublattice otherwise occupied by Ge and Sb. This high intrinsic vacancy concentration is rather unusual. Nevertheless, there are also other semiconductors, such as defective chalcopyrites, which are characterized by a high concentration of intrinsic vacancies. Density functional calculations have been conducted to further investigate the role of intrinsic vacancies. Removing Ge atoms from a randomly occupied cell representing $Ge_2Sb_2Te_4$ and disposing the Ge atoms in a chemical reservoir of elemental germanium was shown to lower the energy of the system. The reason for this lowering of the energy is related to the fact that the highest electronic states occupied for $Ge_2Sb_2Te_4$ are antibonding. Hence, emptying these states by removing Ge atoms lowers the energy of the system. In these calculations an energy minimum was found for $Ge_{1.5}Sb_2Te_4$. Experimentally, however, the resulting crystalline systems are described by the following formula, which relates the number of p-electrons per lattice site, N_p, to the stoichiometry (with n_i giving the number of species i per formula unit):

$$N_p = \frac{2n_{Ge} + 3n_{Sb} + 4n_{Te}}{n_{Ge} + n_{Sb} + n_{Te} + n_V} \tag{2}$$

Those compositions along the GeTe- Sb_2Te_3 pseudobinary line that form stable phases correspond to a number of $N_p = 3$, with $n_V = n_{Te} - (n_{Ge} + n_{Sb})$ (i.e., by balancing the mismatch between the number of anions and cations by intrinsic vacancies). At the same time, this vacancy concentration is required to establish the vacancy layers.

2.1.7 The Significance of Octahedral Arrangements

If one looks at a single structural motif that all crystalline phase change materials have in common, one immediately notices the octahedral-like atomic arrangement. This finding of proximity to an – counting the number of valence electrons – over-coordinated structure (cf. 'hypervalency') and the occurrence of competing Peierls-like distortions, that turns out to be a structural fingerprint of phase-change materials, has important consequences for the electronic structure of phase-change materials as we will see in the following sections. Thus, the quantification of the distortion is an important aspect of structure investigation in the field of phase-change materials. Hence, it is not surprising that many studies have addressed the magnitude of these distortions. The simplest approach is to compare the lattice constant of the rocksalt-type structures with the actual bond lengths. These may for instance be obtained from EXAFS measurements or neutron diffraction data. A split of the six initially equal Ge-Te and Sb-Te bonds (recalling the principal octahedral coordination and atomic alternation) is observed, with distortions of approximately 0.2Å per atom.

2.2 Electronic Structure of the Crystalline Phase

2.2.1 Photoelectron Spectroscopy Studies

Based on the review of the crystalline structure of phase-change materials, we may now turn to the study of the functional behavior that is linked to the electronic structure in the crystalline and amorphous phase. Photoelectron spectroscopy ought to be ideally suited to track down differences for the occupied states. Figure 5 shows XPS-spectra for $(GeTe)_{1-x}(Sb_2Te_3)_x$. Three features are visible in the valence band density of states. The s-like states of tellurium (C), as well as antimony and germanium (B) lie significantly below the p-like states of all three species (A). The latter are located around the Fermi-level and hence are responsible for the bonding.

It is somewhat surprising that the measured spectra for the amorphous and crystalline state look quite similar. This raises the question, how the pronounced optical contrast between both states could be explained. It is not obvious, if the small differences in the valence density of states are sufficient to cause the optical contrast between the amorphous and crystalline state that every user of rewritable optical disks can notice. This task is not only of academic interest, but also has been the motivation of a significant amount of application driven material research. With every new generation of optical storage devices it had to be verified that the chosen phase change materials had sufficient optical contrast at the new wavelength of choice.

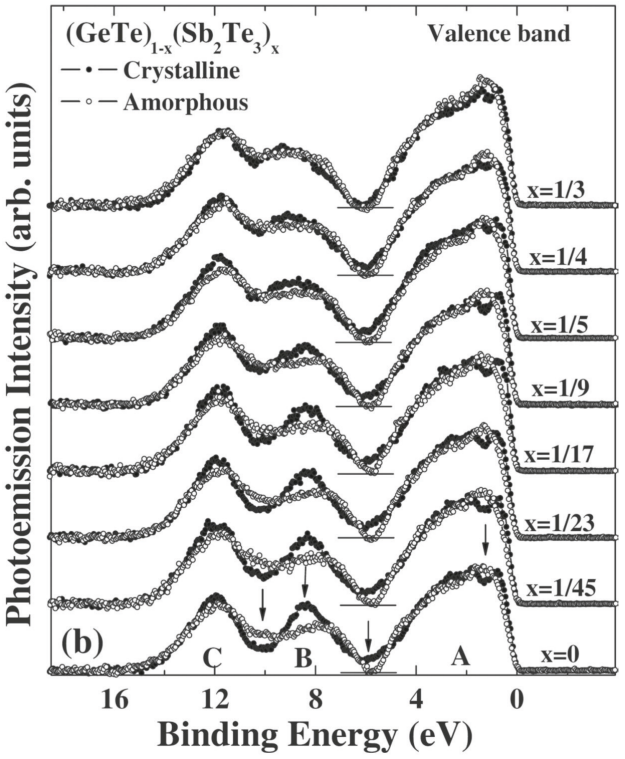


Fig. 5: *The valence band density of states of materials composed as $(GeTe)_{1-x}(Sb_2Te_3)_x$ have been obtained by means of XPS for both the amorphous and crystalline phases. The s-like states (B,C) lie well below the occupied p-like states (A). Only minor differences between the amorphous and the crystalline state can be observed.*

2.2.2 Optical Properties and the Origin of the Optical Contrast

Hence, the optical properties, that also probe the density of unoccupied states, have been investigated by many groups. In the visible range between approximately 1.5 and 3.1 eV, the dielectric function is governed by the interband absorption over the electronic gap. From an extrapolation of the measured spectra it had been extrapolated that in comparison to the amorphous phase, the optical gap is typically smaller, with values of 0.2 to 0.6eV for the crystalline phase. Hence, the gap falls below the typical lower boundary of the energy range that is accessible to ellipsometry. Instead, optical spectroscopy in the infrared can characterize phase change materials. Ideally, such measurements for phase change materials should even be performed significantly below the band gap. This is somewhat surprising, since in this frequency range, phase-change materials should possess a broad transparency window. Nevertheless, the optical properties even below the absorption edge of crystalline and amorphous phase change materials differ significantly. This is shown in Figure 6.

The visual inspection of this figure reveals three differences between the spectra of the amorphous and the crystalline state of phase change materials. The energy range where oscillations, which are indicative for the transparency of the phase change film, are observed is larger for

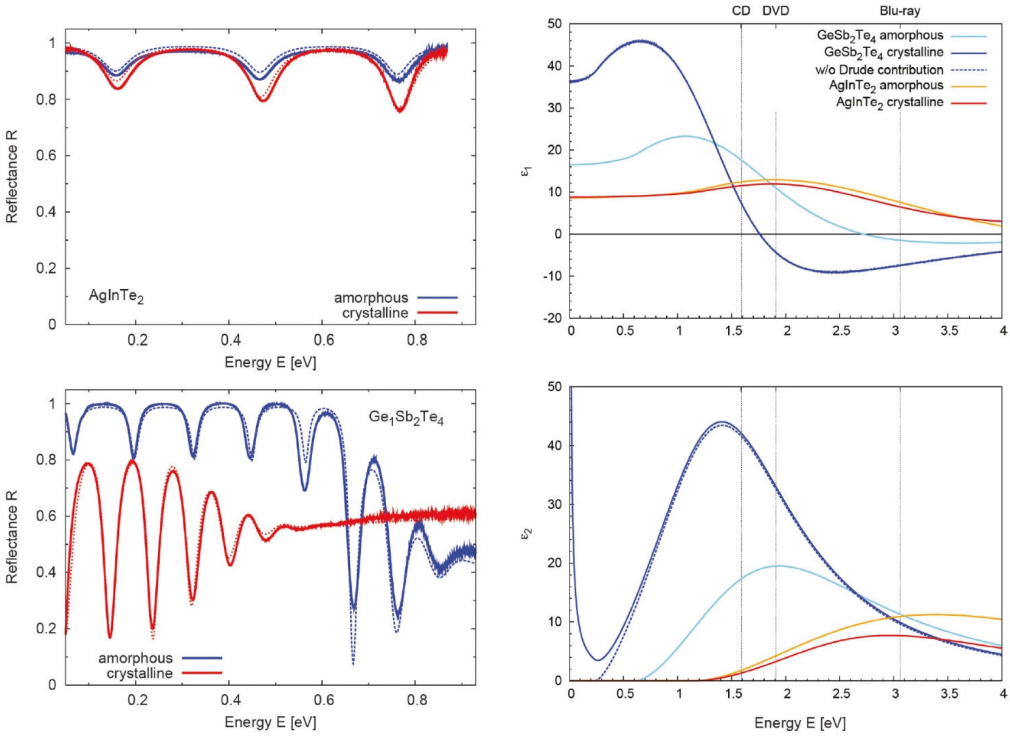


Fig. 6: *A combination of FTIR-measurements (left) and spectroscopic ellipsometry enables the simulation of the dielectric function of narrow-gap materials at and below the optical gap (right). The comparison between a non-phase-change material $AgInTe_2$, and a phase-change material, $Ge_1Sb_2Te_4$, highlights the unique properties of phase-change materials. Only the latter exhibit a significant contrast between the phases. In particular, the gap becomes smaller upon crystallization, and the optical dielectric constant, ϵ_∞, increases significantly.*

the amorphous state. This implies that the amorphous phase has a larger optical gap than the crystalline phase, confirming earlier extrapolations from optical spectroscopy data. In addition, even in the energy range below the band gap, the crystalline films show absorption. This can be seen from the fact that the interference maxima are considerably below unity. Most likely this is due to some collective excitation of charge carriers as will be discussed in more detail below. No such contribution is visible in the spectra of the amorphous phase. The most interesting difference, however, is the different spacing of the reflectance minima. Since in this sample geometry we are performing a simple interference measurement, the spacing of the reflectance minima is governed by the optical thickness, i.e. the product of the film thickness and the corresponding refractive index. The observation that the reflectance minima are much more closely spaced in the crystalline sample cannot be explained by the characteristic 5% decrease in film thickness upon crystallization, which would even lead to a corresponding increase in the spacing of the minima. Instead, the refractive index of the crystalline state has to be significantly higher than the refractive index of the amorphous phase. This distinctive difference can explain the pronounced optical contrast that characterizes phase-change materials. These observations are reflected by the dielectric functions fitted to these spectra. Hence, these or similar measurements are ideally suited to identify and even optimize phase-change materials.

2.2.3 The Stoichiometry-dependence of the Optical Contrast

How does this finding relate to the structure? In Figure 7, the atomic arrangement for a number of chalcogenides is displayed as a function of the average number of p-electrons per atom. This viewgraph shows that for an average number of 2 p-electrons per atom ($N_{sp} = 4$) a tetrahedral atomic arrangement is favorable, while for a larger number of p-electrons this atomic arrangement is destabilized with respect to an octahedral atomic arrangement. Filling a sp^3-bonded system with more than 4 valence electrons requires the occupation of an antibonding orbital, thus the octahedral arrangement, where 5 valence electrons, and hence 3 p-electrons, can occupy bonding states is energetically favorable. All octahedral-like chalcogenides that have been measured so far show a high electronic polarizability in the crystalline state. This raises the question how the octahedral atomic arrangement that is promoted by the p-electrons leads to the high electronic polarizability.

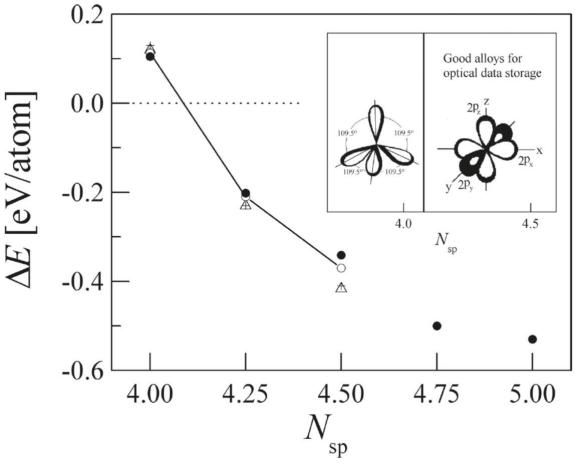


Fig. 7: *Te-containing materials with different average numbers of valence electrons have been studied by DFT-calculations. For each system, the energy difference between four-fold coordinated (chalcopyrite) and six-fold coordinated (rocksalt) structure were calculated. Materials with 4.25 and more electrons were found to form rocksalt structures. Reproduced from ref. [3].*

2.2.4 Resonant Bonding

To understand the high electronic polarizabilty, this finding needs to be related to the electronic structure and atomic arrangements of these materials. We recall that there are about three p-electrons per site available to form covalent bonds in phase-change materials. Save for the atomic distortions, which will not be considered for the moment, the atoms are octahedrally coordinated. Thus, six covalent bonds are established, but there is an insufficient number of electrons to saturate these. This situation is very similar to the case of benzene. It is called resonant bonding (or resonance bonding) and is visualized in Figure 8. To elaborate the importance of this feature for phase-change materials, we shall discuss it in more detail.

Given that a suitable basis to expand the electronic wavefunction, Ψ, consists of saturated bond configurations, ϕ_i, then for a linear chain as shown in Figure 8, it may be expanded as

$$\Psi = \frac{1}{\sqrt{1+\alpha^2}}(\Phi_1 + \alpha\Phi_2). \tag{3}$$

Since all basis configurations ϕ_i are energetically equivalent due to symmetry,

$$\langle\Phi_1|H|\Phi_1\rangle = \langle\Phi_2|H|\Phi_2\rangle = E_0 \tag{4}$$

that is in resonance, the term *resonant bonding* was coined. The total energy is lowered by the resonance energy E_{12},

$$E_{12} = \langle\Phi_1|H|\Phi_2\rangle \tag{5}$$

Resonant bonding leads to a pronounced coupling of the electronic configuration to distortions via α. Thus, the occurrence of resonant bonding is accompanied by a pronounced response of the system to atomic movements or electric fields, for instance. As a result, anomalously large Born effective charges, Z_T, and dielectric constants, ϵ_∞, prevail. The finding of large values of ϵ_∞ in phase-change materials serves as one proof of the occurrence of resonant bonding. Closely related are the rather small bandgaps – generally, ϵ_∞ increases with decreasing bandgap. However, a small gap alone is not sufficient, as in addition also the matrix elements of the optical transition (cf. Fermi's Golden Rule) must be large to produce the significant electronic polarizability enhancement observed in phase-change materials. The works show that structures that support resonant bonding, in particular the crystalline phase of phase-change materials, lead to such high transition matrix elements. Other structures (orthorhombic GeS-type structure, spinel,…) do not give rise to resonance effects since they lack resonant bonding or likewise the alignment of p-orbitals.

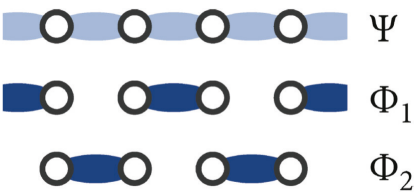


Fig. 8: *The essence of resonant bonding can be visualized using a simple linear chain of equal atoms. The strength of the shading of the covalent bonds symbolizes their saturation. The electronic wave function of the undistorted system, Ψ (top), may be expanded in a basis consisting only of saturated bond configurations Φ_i. The latter are energetically equal due to symmetry that is "in resonance". Thus, the situation drawn at the top corresponds to a superposition of these states. This electronic configuration is easily distorted by external perturbations. Reproduced from ref. [2].*

Resonance effects are a generic fingerprint of crystalline phase-change materials and the corner stone of the optical contrast employed in phase-change devices. These effects, however, are reduced by Peierls-like distortions. Hence, resonant bonding is endangered by these atomic displacements. From the point of view of orbital alignment, it is clear that distortions lead to misalignment. In terms of the expansion into saturated bond configurations presented above, a static atomic distortion favors one configuration over the other and thus counteracts the resonance. Nevertheless, for small distortions resonance effects are weakened but prevail; density functional perturbation theory calculations on GeTe prove, that, while the Peierls-like distortion and the subsequent cell distortion significantly reduce the values of Born effective charge and optical dielectric tensors, they nevertheless remain anomalously large. Therefore, it is proposed that the search for phase-change materials may be directed to those materials that exhibit resonant bonding and only a limited level of distortion.

2.2.5 Stoichiometry-dependence of the Occurence of Resonance

In order to assist this search, it is desirable to have a simple theoretical scheme that enables a prediction on whether a material may be expected to exhibit the desirable characteristics prior

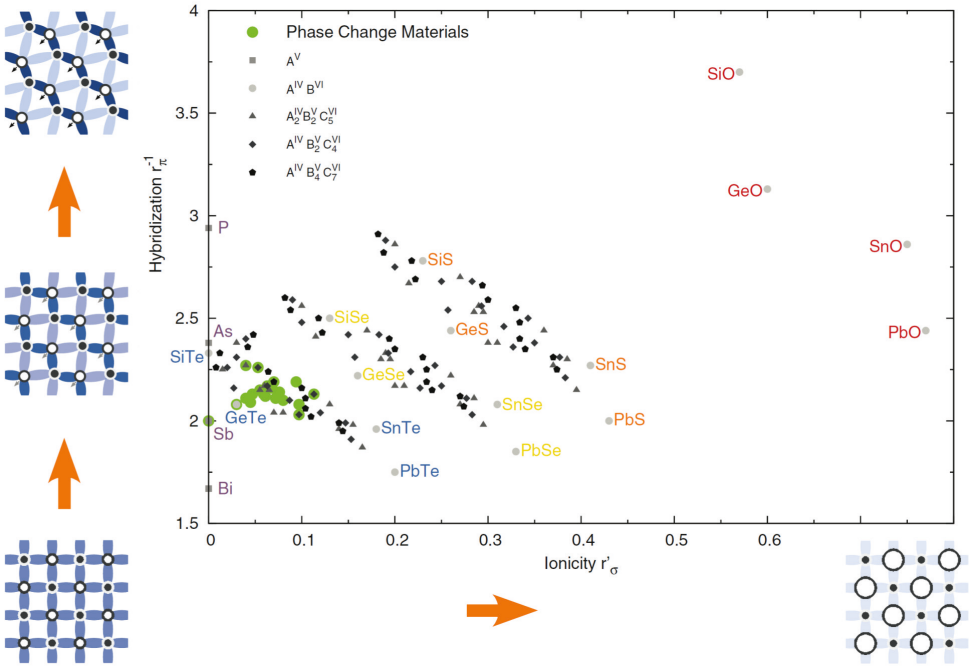


Fig. 9: *Empiric map for materials with about three p-electrons per atomic site and even numbers of anions and cations. The axes that span the map are the tendency towards hybridization, r_π^{-1}, and the ionicity, $r_{\sigma'}$, both defined in the text. The coordinates of a large number of materials have been calculated. Phase-change materials are located within a small region of the map that is prone to the occurrence of resonant bonding. The graphs on the outside illustrate the weakening of resonance effects as one leaves this region due to the formation of less, more saturated covalent bonds via distortions or due to charge localization at the ions due to increasing ionicity. Reproduced from ref. [4].*

to characterization. Such a scheme has recently been proposed [4]. Based on the work by Littlewood [5], a two-dimensional map is constructed. It is spanned by two coordinates, where the sums run over the anions and cations, respectively. n_i denotes the concentration of species i, while $r_{p,i}$ and $r_{s,i}$ refer to the radii of p- and s-orbitals, respectively. These radii are obtained from pseudopotential calculations. The first coordinate, r'_σ, provides a measure of ionicity. The second, r_π^{-1}, provides a measure of the tendency towards hybridization. For materials with approximately even numbers of anions and cations as well as $N_p = 3$, the resulting map is shown in Figure 9. The impact of both coordinates on the structures is shown by the schematic graphs at the border of this figure. An increasing hybridization favors distortions, which counteract the resonance character of the bonding. Increasing ionicity weakens the covalent (resonant) bonds at the expense of charge localization at the ion cores, and hence decreases the signature effects of resonance bonding. Thus, for resonant bonding with small distortions to occur, it is expected that only materials in the lower left corner have the potential to be employed as phase-change materials. Indeed, phase-change materials (marked in green) that have been identified empirically are found in this region. This underlines the potential of this simple scheme. However, it also shows that there is little room for finding better materials (with simple stoichiometries) than the already known ones. Future research may therefore attempt to transfer this scheme to other stoichiometries and other values of N_p. In that sense, the present two-dimensional map is only one projection plane within a higher-dimensional composition space. Though, it is a particularly important one, since it hosts typical phase-change materials including antimony, GeTe and the Ge:Sb:Te-class as well as materials derived by isoelectronic exchange.

It has to be stressed that careful application of the map concept is advised. For instance, it is meaningless to put materials with an average number of p-electrons of two or four, or a pronounced deviation in composition from an equal amount of anions and cations such as e.g. $SnSe_2$ onto the same projection plane. Instead, a different projection plane should be chosen. Elements with d-states close to the Fermi-level are also not incorporated in the present scheme. Finally, the stability of a composition against phase separation also cannot be inferred from a map. Further information on the development of structure maps may for instance be obtained from the work of Pettifor.

To this point, we have identified the unique structure and bonding in the crystalline state as well as the resulting properties. From the studies of the electronic structure and the optical properties, we know already that the former is apparently similar in both phases, yet the latter is very different, enabling the use of phase-change materials in optical storage devices. Thus, along the lines of Sir Francis Crick, the question arises what structure the glassy phase exhibits in order to understand why resonance effects are limited to the crystalline phase.

2.3 Atomic Structure of the Melt

Since amorphization in phase-change devices proceeds via melt-quenching, it is clear that properties of the melt, such as its structure, are important for the understanding of phase-change recording. This is not only true since electrothermal modeling requires these properties as input, but also glass formation and structural properties of the glass must be expected to reflect properties of the liquid phase, since to a first approximation, a glass resembles a frozen-in liquid configuration. Therefore, it is evident that investigations of the liquid phase are important for the understanding of the amorphous phase and the kinetics of the phase transformation. To facilitate this, mainly molecular dynamics simulations as well as x-ray and neutron diffraction have been employed so far. Since the required elevated temperatures pose severe experimental challenges, the main focus is commonly on the structure.

The general result of these investigations, that are to be addressed in the following, is that materials successfully employed as phase-change materials are characterized by an octahedral bonding geometry of the atoms, with bond angles ranging around 90 degrees. Related chalcogenides which do not show phase change properties, on the contrary, deviate from this scheme, exhibiting tetrahedral sites with angles in the range of 109.5 degrees as expected for sp^3-hybridization. This finding has recently been obtained from neutron diffraction experiments performed on a wide range of phase-change materials and related materials. These investigations revealed that only phase-change materials exhibited octahedral liquids. Moreover, the average number of valence electrons per atom was identified as an order parameter, with the octahedral-to-tetrahedral transition occurring at a threshold of about 4.25. So far it seems as if the octahedral-like atomic arrangement is the most important fingerprint of the structure of liquid phase-change materials. This is important since in recent years the identification of octahedral or tetrahedral bonding geometries in the amorphous and to a lesser extent the liquid phase has been a recurring motif in the literature.

2.4 Atomic Structure of the Amorphous Phase

The structure of the solid amorphous phases of phase-change materials has been studied primarily by means of x-ray techniques that are sensitive to local atomic environments, e.g. extended x-ray absorption fine structure (EXAFS) measurements, neutron diffraction and molecular dynamics-simulations. With the amorphous phase of phase-change materials generally representing a covalently bonded network, the aim is to reveal the coordination number and the bonding geometry of each atomic species. This allows to assess the amorphous phase in terms of the previously introduced concepts and aids the understanding of the transition kinetics on an atomistic level.

Since most effort was put into research on prototype Ge:Sb:Te materials, we shall focus on these results. From EXAFS measurements on GeTe and $Ge_2Sb_2Te_5$, it was found that chemical ordering takes place in the amorphous phase, leaving Te-atoms with germanium (and antimony) neighbors. The Ge-Te bond length was determined to be around 2.6 Å and the coordination number of Ge was derived as four. Since the bonds could not be angular resolved, tetrahedrally coordinated germanium due to sp^3-hybridization in $GeTe_4$-configuration was assumed.

More detailed structural investigations have focused on resolving limitations of the x-ray-based studies. The combination with neutron diffraction data yields better contrast between Sb- and Te-atoms, that can hardly be distinguished by x-rays alone. The use of Reverse Monte Carlo method to simultaneously fit all data sets enables the simulation of the amorphous structure in large models, which allow for more than just a few local motifs and can statistically be evaluated. Employing this approach for as deposited-amorphous samples of $Ge_2Sb_2Te_5$ and $GeSb_2Te_4$, chemical ordering was confirmed. In particular, homopolar Te-Te- and Sb-Sb-bonds were ruled out, while a significant portion of Ge-Sb and Ge-Ge bonds was reported. Those were referred to as 'wrong bonds' as they were not included in the original model. Coordination numbers were found to closely match the 8 - *N*-rule. A predominance of the suggested $GeTe_4$-tetrahedra was not obtained, leading the authors to reject approximations of the structure of the amorphous phase using simple local motifs. However, while the bond angle distribution for Te-Sb-Te showed a preferentially octahedral environment, the bond angle distribution for Te-Ge-Te indicated higher bond angles that are rather in line with tetrahedral environments. The latter finding is in contrast to the expectation we infer from studies on the liquid phase, that indicate an octahedral structure also for germanium, but also other studies of the amorphous phase.

Due to the obvious differences between structure models derived from experiments alone, various large-scale molecular-dynamics simulations have been performed in recent years, enabling an unprecedented and detailed insight into structure and properties of the amorphous phase of phase-change materials. The general result of these studies is a confirmation of chemical ordering, leading to the occurrence of 4-rings with pronounced AB-alternation. However, also a significant number of homopolar bonds is obtained. The most important conclusion of these studies is a predominance of octahedrally coordinated atoms, while those atoms that show a deviating tetrahedral atomic arrangement often form homopolar Ge-Ge bond configurations.

We are now in a position to compare the atomic arrangement and the resulting electronic structure in the amorphous, liquid and crystalline state of phase-change materials. All phases are characterized by a predominance of octahedral coordination. This implies that bonding occurs mainly via p-electrons. Furthermore, there is an alternation between anionic and cationic species ('AB-alternation'). The structural similarities are reflected by the electronic structure. In this situation, where the phases between which switching is facilitated are so similar, the question arises what the pronounced property contrast between them stems from. Yet, the answer was already given in Section 2.2.4, when we noted that the crystalline phase of materials that are successfully employed as phase-change materials are close to resonance conditions. This, however, is only feasible if distortions are small and medium-range order prevails. Hence, it is the crystalline phase that sets phase-change materials apart, at least as far as the property contrast is concerned.

3 Material Properties

In the preceding sections, a concept was developed which relates structure to bonding and the resulting optical properties. Now, our aim is to extend this framework to assess vibrational and thermal properties. Then, we review the current state of understanding of the electrical properties that these materials are characterized by.

3.1 Vibrational and Thermal Properties

The lattice dynamics of the crystalline phase of phase-change materials have primarily been assessed by Raman-spectroscopy and density functional theory calculations. In addition, insight into the vibrational properties has been gained by analyzing Debye-Waller factors from structural investigations. Since Debye-Waller factors, B, serve as a measure of the 'smearing' of atomic positions, they comprise two effects. On the one hand, the Peierls-distortion is typically subsumed as a static contribution. This approach is meaningful if isotropic atomic displacements from high symmetry-positions are assumed. On the other hand, a dynamical, temperature-dependent component reflects the thermal excitation of atomic vibrations. In the harmonic approximation, $B(T)$ factors are expected to exhibit a linear increase at elevated temperatures (temperature-derivative of the Bose-statistics). Debye-Waller factors have, for instance, been obtained by means of density functional-theory calculations and x-ray diffraction measurements. In the latter work, even a super-linear increase at high temperatures has been obtained. Generally, the increase in the crystalline phase is larger than the one in the amorphous phase. This finding has been designated as a signature of phase-change materials, that is associated with the anharmonicity of the potential seen by the atoms due to the Peierls-like distortions (see figure 10).

The vibronic properties of amorphous materials have been investigated for various reasons; experimentally, the observation of certain modes, for instance in Raman spectroscopy, may serve

as fingerprint of the presence of corresponding local structural motifs such as $GeTe_4$-tetrahedra. However, a possible shifting of the frequencies of such modes, for instance due to the environment of a local motif, leaves room for misinterpretation. Thus, the support by computer simulations proves useful. For instance, based on the analysis of the vibrational density of states projected onto different types of germanium atoms, tetrahedrally coordinated and homopolarly bonded germanium could be identified to be responsible for the energetically highest vibrational excitations in $Ge_2Sb_2Te_5$ and in GeTe. Furthermore, merely as a side product of computationally expensive molecular dynamics simulations, also information on atomic diffusion constants (viscosity) is provided at certain temperatures given sufficient integration times.

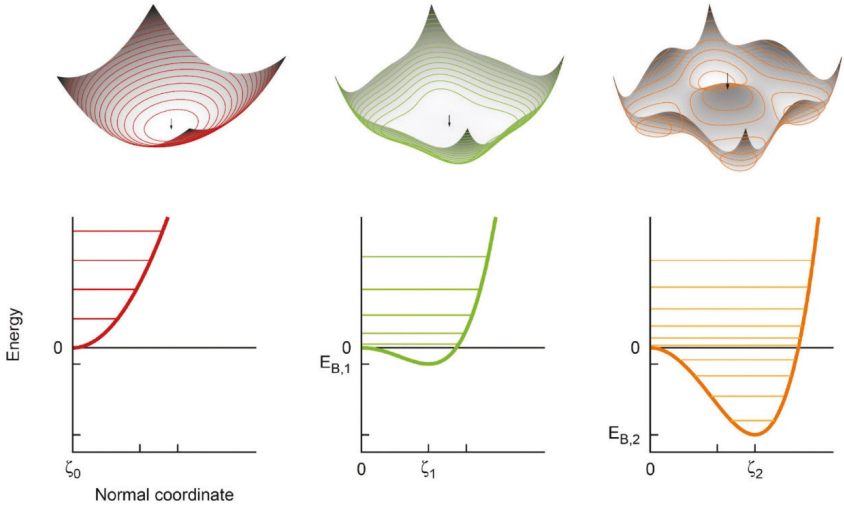


Fig. 10: *Peierls-like distortions lead to a peculiar shape of the energy landscape, i.e. the energy as a function of the atomic positions (in terms of a normal coordinate). From left to right, the undistorted case, a small, and a strong distortion are considered. The energy levels are indicated in the bottom row by horizontal lines. In the two border-cases the atoms only probe a rather harmonic potential at small temperatures. In the intermediate case, however, a pronounced anharmonicity affects the vibrational and thermal properties. Reproduced from ref. [2].*

The thermal properties of phase-change materials are of interest mainly for two reasons; as far as devices are concerned, knowledge about the heat propagation is of importance in order to ensure low power consumption by heat confinement. It is also interesting to note that phase-change materials are very similar to thermoelectric materials. Bi_2Te_3 and PbTe, for example, are well established thermoelectrics, differing from Sb_2Te_3 and GeTe only by isoelectronic exchange. The figure of merit of thermoelectrics, Z, is defined as

$$Z = \frac{\sigma S^2}{\kappa}, \tag{6}$$

where σ is the electrical and κ the thermal conductivity, S is the Seebeck coefficient. A suitable material is characterized by a high electrical conductivity as well as small thermal conductivity. Given our previous argument on the energy landscape of the crystalline phase, it is clear that anharmonicity due to Peierls-distortions may serve as such an intrinsic mechanism suitable to limit the lattice contribution to the thermal transport. It is expected to be pronounced in materials further

down in the map, so it is little surprise that PbTe for instance exhibits the desired properties. The map concept may thus be helpful to obtain materials suitable for thermoelectric applications.

3.2 Electrical Properties

In recent years, the electrical properties of phase-change materials have attracted considerable interest as research shifted from optical to electrical memories. Consequently, the conductivity in both the crystalline and the amorphous state shall be briefly reviewed here. The fact that it differs significantly between the two phases, as the exemplary measurement of the resistivity in Figure 11 shows, allows to distinguish between logical states. In $Ge_2Sb_2Te_5$, for instance, the specific electrical conductivity at room temperature changes upon crystallization from $\sim 4 \times 10^{-3}\ \Omega^{-1}cm^{-1}$ to $\sim 1.5 \times 10^{3}\ \Omega^{-1}cm^{-1}$, that is by six orders of magnitude. Further analysis shows that the concentration of charge carriers in the crystalline phase is particularly high, reaching values on the order of $\sim 1 \times 10^{20}\ cm^{-3}$. This is not a consequence of the small gap and thermal excitation, but induced by a pronounced shift of the Fermi-level towards or even into the valence bands due to defects. In anticipation of the results presented in this section, this indicates that the states around the Fermi-level and their nature dominate the electrical behavior of phase-change materials. Figure 12 visualizes the principal situation in Ge:Sb:Te-materials.

3.2.1 Defects and Localization in the Crystalline Phase

For the crystalline phase, the most notable fact is that phase-change materials behave as extrinsic semiconductors. While plain narrow gap semiconductors are expected, the Fermi-level is measured to reside close to or even within the valence band, giving rise to p-type conductivity.

This behavior is observed in a number of experiments. Measurements of the resistivity as a function of temperature show a change of slope from amorphous to the metastable and the stable crystalline phase. The change in the activation energy exceeds the change of the band gap. When even a positive slope upon temperature increase is observed, the transition to a degenerate semiconductor becomes obvious. Due to this Fermi-level shift, free carriers are more easily thermally created in the crystalline state. Low-temperature Hall measurements have been conducted. The Hall-coefficient confirmed p-type conduction in the crystalline phase, and did not exhibit a freeze-out of carriers down to 5K. Thus, for the investigated samples, the Fermi-level must have been within the valence band. Upon the metastable to stable crystalline phase transition, the charge carrier concentration increased by about 30%, but the mobility increase was much more pronounced.

The Fermi-level shift towards/into the valence band in the crystalline phase has been attributed to the occurrence of defects. The case has been extensively studied for GeTe; the calculation of the energy of formation of various defect types and charge states shows that particularly Ge-vacancies are easily formed, requiring little thermal energy. Depending on the position of the Fermi-level, these defects are (negatively) charged or neutral. The presence of significant amounts of vacancies in the germanium-sublattice has been confirmed by different experiments. The same mechanism of defect formation, vacancies on the cation sublattice, is commonly anticipated to hold for other Ge:Sb:Te-based phase-change materials.

In a recent study a transition between thermally activated (non-metallic) and metallic transport has been observed in crystalline $GeSb_2Te_4$ with increasing annealing temperature. The non-metallic behavior has been attributed to a disorder induced localization in the metastable rock-salt-structure. At the transition to the metallic state, the corresponding electronic mean free path only amounts to 0.8nm, which is interpreted as demonstrating a remarkably high level of atomic disorder. Furthermore, these authors suggest that controlling the degree of atomic ordering hence provides a mechanism to adjust the electronic properties.

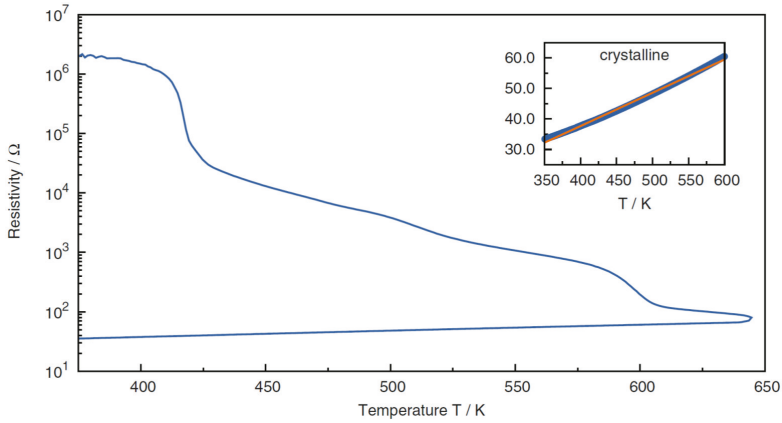


Fig. 11: *Shown are the phase transitions of a sample of $GeSb_2Te_4$ measured by the Van der Pauw-method. Crystallization occurs at around 420 K and leads to a significant drop in resistivity. A second irreversible, solid-solid transition is observed at about 600 K, which is identified as the transition from the metastable rocksalt- to the stable hexagonal phase. The charge transport in the amorphous phase is reasonably well described as thermally activated, while the crystalline phase obtained here exhibits a metal-like temperature-dependence (inset).*

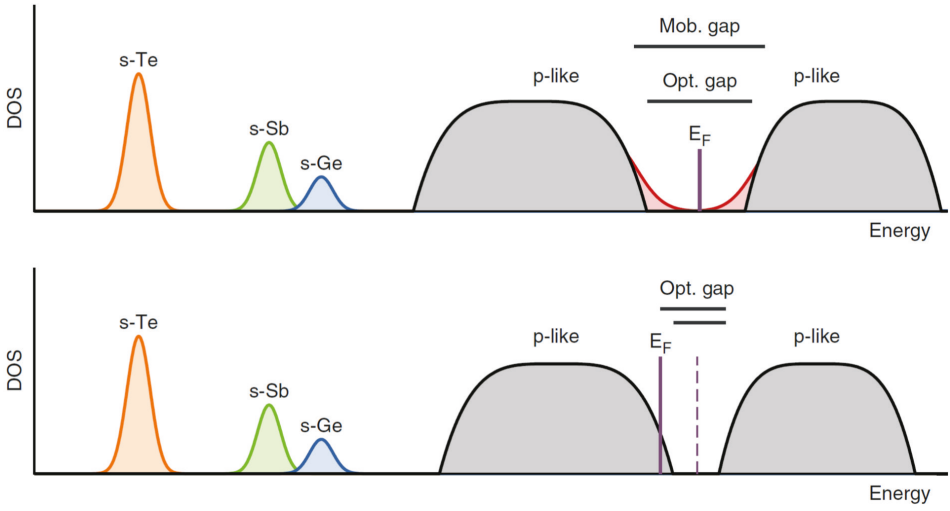


Fig. 12: *The schematic electronic densities of states of Ge:Sb:Te-materials in the amorphous (top) and crystalline (bottom) phase are compared. The s-states of the atoms are well separated below the p-like states. The latter are broadened and overlap. The optical gap of the amorphous phase is larger than in the crystalline state. Localized states (red) due to band tails (or defects) are found within the mobility gap. The Fermi level is pinned within the optical gap. In the crystalline state, the Fermi level is typically shifted towards or even into the valence band due to the formation of defects. Then, the optical gap exceeds the electronic gap due to the Burstein–Moss-effect. Reproduced from ref. [2].*

4 Applications and Outlook

In the last section, our aim is to give an overview over the various applications that employ phase-change materials and their future development.

4.1 Optical Storage

Up to now, optical data storage devices represent the most common application of phase-change based recording. With the introduction of blue laser light and maximization of the numerical aperture, the race for higher resolution and hence the development of this field may at first seem exhausted. Room for increasing storage capacity would be left only by increasing the number of data layers per disk. At present, a 100GB disk employing three data layers marks the upper limit for such data storage devices, paving the way to the fourth generation of commercial optical phase-change based media. However, another way of increasing capacity has been found in the course of phase-change research that is near-field recording. Here, an optical element is placed close the surface of the medium, at a distance that is much smaller than the optical wavelength. This allows to employ evanescent waves to manipulate bits on the medium beyond the diffraction limit. The so-called Super-RENS effect (super-resolution near-field structure) refers to the observation that structures smaller than the optical wavelength can be obtained by combining phase-change films with an additional thin layer in optical near-field range. In other words, the part of the optical system that needs to be located very close to the phase-change film to enable near-field recording is incorporated into the disk structure itself. Given the variety of proposed explanations but also the potential benefit, the field of phase-change based near-field recording remains a potentially rewarding research area.

4.2 Electronic Storage

When phase-change materials were introduced in 1968 by Ovshinsky [1], electronic memories were among their first suggested applications. Nevertheless, only now due to the availability of fast-switching phase-change materials and the ability to create nanoscaled structures, it is possible to create competitive, non-volatile phase-change based electronic memories: *phase-change random access memory*, usually abbreviated *PRAM* or *PCRAM*. As can be seen from the PTE-diagram shown in Figure 13, operation can proceed on the timescale of few nanoseconds which is orders of magnitude faster than Flash. This puts PCRAM in a position of a universal memory, which combines the best of both DRAM and Flash. Moreover, the attainable scalability even

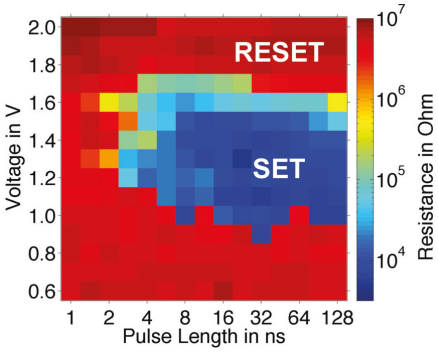


Fig. 13: *Characterization of a PCRAM-device employing GeTe as the active material using an electric tester. Crystallization, the time limiting process involved in phase-change recording, can be triggered by applying pulses as short as only about 5 ns.*

surpasses the existing memories, while having low power consumption. Hence, it is not surprising that the development of such memory cells has benefited from many industrial contributions.

Two principle designs of PCRAM cells have been proposed, line-cells and vertical structures, cf. Figure 14. A line-cell is simply a lateral line of a phase-change material that connects two electrodes. Given the large occupied area, its use is mainly limited to research projects – favoring the possibility to access the cell from the top – rather than actual memory devices. Though, some applications may exploit the ease of fabrication and low power consumption of these cells. The other type is a stack of layers, where a thin volume of phase-change is on top of a highly resistive heater element. Here, a portion of the phase-change volume that is close to the heater, which typically features a reduced diameter to increase the current density, is switched between the phases. Recently, an improved version of a PCRAM-cell has been realized that integrates the cell selector into the cell design. By the use of an Ovonic threshold-switch (OTS), a selector with no more than the same spatial footprint as the memory cell itself (i.e., $4F^2$, F being the feature size) is possible. This is to be contrasted with other, spatially more demanding designs mentioned in a recent review. The OTS is just a thin layer of a material that exhibits threshold-switching. If the voltage drop over the OTS due to the voltage applied between word-

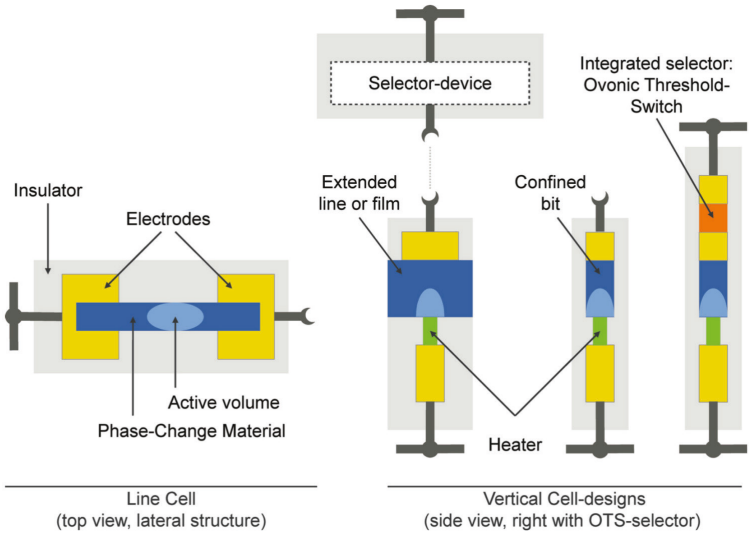


Fig. 14: *Comparison of typical designs of phase-change cells. While the lateral design is of particular interest to researchers, the vertical structures have a smaller lateral size and are thus favorable to achieve high storage densities. Here, one may distinguish between two subtypes with extended or confined volumes of a phase-change material. The former employ a thin layer (or line) of a phase-change material, while the latter involve pores in which the active volume more closely matches the actual volume of the material. Though this allows for a better control of the thermal environment of the cell, it also necessitates advanced production techniques that allow for high aspect ratios. Since the cells typically require also some selector devices, their actual spatial footprint is unfavorably enlarged. Thus, the concept shown on the very right is promising, where the selector is integrated into the cell design, not requiring any additional lateral space. It may consist of a material that exhibits threshold-switching and does not change its characteristics–for instance due to crystallization–throughout operation. Reproduced from ref. [2].*

and bit line exceeds its threshold field, the memory cell, which is in series with the OTS, is selected. The simple design may also allow the stacking of layers of PCRAM-cells, increasing the storage density by making use of the third dimension.

PCRAM makes extensive use of the threshold switching effect. In order to supply sufficient Joule heating power $P_J = U \cdot I = U^2/R(U)$ to quickly raise the temperature to levels high enough for crystallization to occur on a short timescale (cf. Figure 13), very high voltages would be required, if threshold switching would not occur. Thus, it allows to avoid voltage upconversion and the problems linked with high voltages in nanoscaled-electronics.

So far, we have described the advantages of PCRAM over competing memory technologies in terms of its non-volatility, speed and attainable storage density. Another aspect, however, is cyclability (i.e., the number of possible write-cycles). The two main wear processes identified so far are electromigration and void formation. It is found that for cells based on Ge:Sb:Te-materials, the spatial distribution of the elements changes upon multiple set and reset operations. In particular, antimony accumulates at the cathode, pushing germanium aside. Thus, the composition in the active volume and thereby the cell properties change. Operation at reversed polarity, though, can 'repair' such a cell. The density change upon crystallization and atomic mobility may also hamper the electrical contact via void formation. In addition, degradation of the electrodes (e.g., diffusion into the phase-change material) and phase segregation in the case of non-stoichiometric materials may also be regarded as limiting factors. Nevertheless, though the exact number of possible cycles depends on a variety of factors, it has been proven to exceed the corresponding value of Flash by several orders of magnitude.

4.3 Other Applications

Beyond the commercialized optical and electrical memories introduced before, a variety of possible future uses and research opportunities for phase-change materials have been investigated. The concept of probe-based storage envisions the use of an array of conductive AFM-tips to parallely switch bits of a phase-change layer. It combines high resolution (i.e., storage density) with 'simple' media. The interest in one-dimensional systems has been adopted by few phase-change researchers that have successfully synthesized and characterized phase-change nanowires for data storage.

Other authors seek to combine the reversible amorphous-to-crystalline transition with other material properties or device characteristics, which enable additional degrees of freedom to store or access data. The possibility of polarity-dependent resistance switching has been investigated by employing Sb-excess $Ge_2Sb_2Te_5$ in electrical cells, where additional conductive Sb-filaments can be formed or broken. Thus, there is a second mechanism to set the resistivity of a cell. Song et al. incorporated iron into phase-change materials. This way, ferromagnetic, so-called *phase-change magnetic materials* could be obtained. For sufficiently small Fe-concentrations, it was possible to control the magnetization by switching between the phases. This observation was attributed to the difference in carrier concentration between both phases, since the carriers were supposed to provide the required, indirect interaction between Fe-precipitates.

Finally, the fact that the resistivity of a phase-change cell depends on its history (i.e., more than two logical states can be represented by one cell, so that multi-level storage becomes feasible) has led to the proposal of using such devices to emulate the behavior of synapses, paving the way for cognitive information processing. In that sense, phase-change based data storage does not only hold the potential to serve as a fast and reliable, universal non-volatile memory. Moreover, this technique could also revolutionize the way we process data.

Acknowledgements

We gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft within SFB 917 (Nanoswitches) as well as the ERC through an Advanced Grant. Furthermore, I would like to thank the members of the phase-change research group at RWTH Aachen University for fruitful discussions.

References

[1] This manuscript is based on the review by D. Lencer, M. Salinga, and M. Wuttig, *Advanced Materials* 2008, 7, 972

[2] D. Lencer, *Design Rules, Local Structure and Lattice-Dynamics of Phase-Change Materials for Data Storage Applications*, Ph.D. thesis, RWTH Aachen University, 2010.

[3] S. R. Ovshinsky, *Physical Review Letters* 1968, *21*, 1450.

[4] D. Lencer, M. Salinga, B. Grabowski, T. Hickel, J. Neugebauer, M. Wuttig, *Nature Materials* 2008, 7, 972.

[5] P. B. Littlewood, *CRC Critical Rev. Solid State Mater. Sci.* 1984, 11, 229.

Other references

[6] M. Wuttig, D. Lusebrink, D. Wamwangi, W. Welnic, M. Gillessen, R. Dronskowski, *Nature Materials* 2007, *6*, 122.

[7] M. Wuttig, N. Yamada, *Nature Materials* 2007, *6*, 824.

[8] B. Huang, J. Robertson, *Physical Review B* 2010, *81*, 081204.

[9] K. Shportko, S. Kremers, M. Woda, D. Lencer, J. Robertson, M. Wuttig, *Nature Materials* 2008, *7*, 653.

[10] C. Steimer, V. Coulet, W. Welnic, H. Dieker, R. Detemple, C. Bichara, B. Beuneu, J. P. Gaspard, M. Wuttig, *Advanced Materials* 2008, *20*, 4535.

[11] T. Siegrist, P. Jost, H. Volker, M. Woda, P. Merkelbach, C. Schlockermann, M. Wuttig, *Nature Materials* 2011, *10*, 202.

D 7 Threshold and Memory Switching Kinetics of Phase Change Materials

M. Salinga
Institute of Physics (IA)
RWTH Aachen University

Contents

1 Introduction **2**

2 Crystallization **4**
2.1 Growth of a crystallite 5
2.2 Nucleation 6
2.3 Experimental quantification 7

3 Threshold Switching **11**

4 Experimental observations in memory devices **13**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

Phase Change Materials owe their name to the ability to rapidly change their structural configuration between otherwise stable states when experiencing according excitation. Together with the pronounced contrast in electrical resistivity between the different configurations this builds the foundation for an information storage technology called “phase change memory”. These materials are not only stable in crystalline phases, they can also exist for very long times in amorphous solid states without undergoing a phase transition into energetically more favourable crystalline structures.

The utilized process for creating an amorphous solid from a crystal is melting the material in order to destroy the atomic long range order and then cooling it down so quickly that the atoms have not enough time to find their way back into the energetically preferred crystalline phase. This way the atoms get stuck in their disordered configurations. At lower temperatures (below 100°C) the atomic mobility is so low that even after years the amorphous material does not crystallize. However, if the disordered material is brought to elevated temperatures, crystallization can take place within few nanoseconds.

This ability to crystallize with enormous speeds in turn poses a strong requirement to the cooling rates during melt-quenching. In order to reach the necessary cooling rates in the range of 10^9 -10^{11} K/s the heat must be able to leave the material very effectively. Thus, the phase change material must be in good thermal contact with materials of high heat conductivity and it must be structured in a way so that its surface-to-volume ratio is high.

These requirements are intrinsically fulfilled by modern memory technology, where phase change materials are not only in electrical and thermal contact with metal electrodes. The push for ever higher data densities leads also towards smaller and smaller structures on the nanometer scale implying individual memory cells with a very large surface-to-volume ratio.

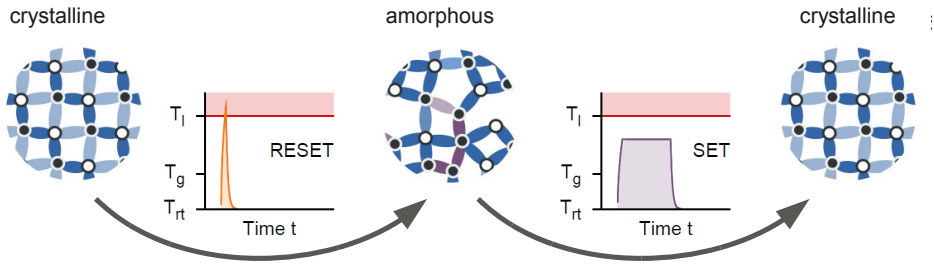


Fig. 1: *The operation principle of phase-change devices is based on the reversible switching between the crystalline and amorphous state. Amorphization (also called RESET - operation) of a bit proceeds via melt-quenching, employing short current pulses as heat sources. The resulting huge temperature difference between the confined melt and the surrounding material leads to extremely high cooling rates. Thus, the disorder of the liquid is frozen in. Crystallization (SET-operation) requires annealing of an amorphous bit at a temperature below the melting temperature for the atoms to adopt the energetically favorable crystalline order. From [3] and [4].*

Accordingly, the switching kinetics in the amorphization process, the so-called RESET, is dominated by the thermal surrounding of a phase change memory cell and much less influenced by fundamental physical properties of the phase change material itself.

Because the thermal environment has certainly a big influence on the temperature distributions that can be realized in a phase change memory cell, it is also important for the recrystallization process, often called SET, where the amorphous material is heated into a temperature regime of fast crystallization. However, the SET speed can be significantly limited by the intrinsic crystallization kinetics of the phase change material used. An optimization of the opposing demands of fast SET speed, i.e. crystallization, and long retention times, i.e. stability against crystallization, is a central task in the research on phase change materials for memory applications. Thus, crystallization kinetics of phase change materials will be discussed in more detail in Chapter 2.

While the structural phase transitions are controlled by temperature and the thermal environment has influence on the temperature profiles that can be realized, elevating the temperature inside a phase change memory cells needs a heat source. In electronic memories based on phase change materials the heat is produced by Joule heating, i.e. by sending an electrical current with significant current density through a material with finite resistivity. Irrespective of whether Joule heat is dominantly created in the phase change material itself or the neighbouring electrode material contributes too, the phase change material must allow significant currents to flow from one electrode to the other for the device temperature to reach the regime of fast crystallization (> 600 K). While in their crystalline states phase change materials are highly conductive, the large resistivity of their amorphous states prohibits large current densities. It is the strong non-linear increase of the conductivity at higher electrical fields and in particular a phenomenon called threshold switching that allows for sufficiently high current densities. The details of this unconventional electrical excitability of the amorphous states of phase change materials will be presented in Chapter 3.

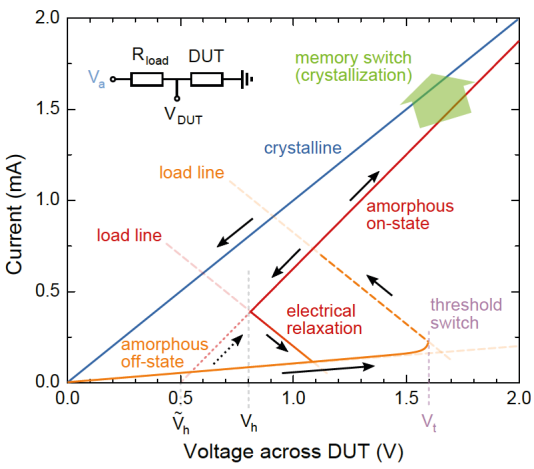


Fig. 2: *Illustration of a typical current-voltage characteristic curve for phase change materials. In the amorphous off-state (orange) the conductivity is very low (in reality even 100 times lower than illustrated). At the threshold voltage V_t the conductivity of the amorphous phase change material increases rapidly by orders of magnitude without a phase transition to the crystalline state. The material is now in the amorphous on-state (red). If the applied voltage in the amorphous on-state is increased until the temperature in the phase change material exceeds the crystallization temperature, the memory switch to the crystalline phase (blue) with a high conductivity occurs. From [3] and [5].*

2 Crystallization

Crystallization in phase-change materials has been successfully described by the interplay of atomic mobility and driving force for crystallization (see Figure 3). For temperatures close to the glass transition temperature T_g the driving force $\Delta G(T)$ is large, while the atomic mobility is low hindering the crystallization process. This way, the small mobility is the reason for high data retention in phase-change memories at low temperatures, i.e. ambient/operating temperature of the device. In this regime the slow processes of nucleation and growth are easily accessible experimentally e.g. by transmission electron or atomic force microscopy. At temperatures close to the liquidus temperature T_l the driving force vanishes. Therefore, the crystallization is also suppressed very close to the melting point despite the fact that the atomic mobility is extremely high. The technologically relevant fast switching in phase-change memories can be realized in the regime of fast crystallization between T_g and T_l, where the atomic mobility is very high and the driving force still significant.

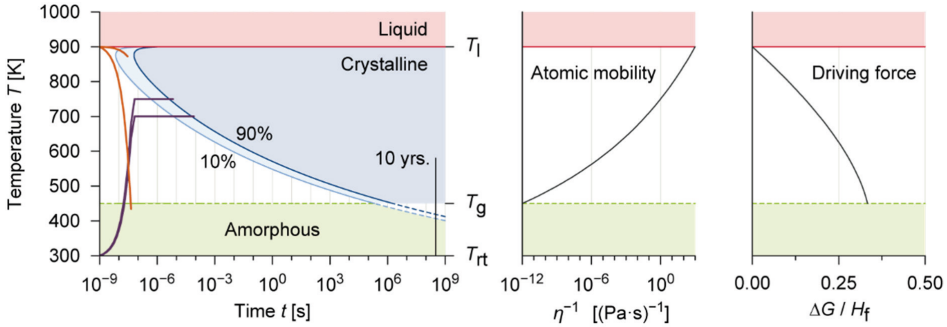


Fig. 3: *Temperature dependencies of crystallization kinetics (left), atomic mobility (middle) and driving force for crystallization (right). The atomic mobility is linked to the viscosity via the Stokes-Einstein equation in an inversely proportional way. For increasing temperature the atomic mobility increases while the driving force for crystallization, the difference in Gibbs free energy ΔG between the liquid and crystal, decreases to zero at the liquidus temperature T_l (in this example at around 900 K). ΔG is plotted in units of the heat of fusion H_f. Although the driving force is large at low temperatures close to the glass transition temperature T_g, the atomic mobility is strongly reduced. Therefore, the regime of fast crystallization can be found at intermediate temperatures between T_l and T_g. Due to the interplay of the atomic mobility and the driving force, the crystallization does not occur immediately, when cooling a liquid below T_l. For rapid cooling rates of about 10^{10} K/s (orange line) the regime of fast crystallization can be bypassed and the material ends up in the melt-quenched amorphous phase. In case of much slower cooling rates, the phase-change material crystallizes. The blue curves indicate how much time it takes at a certain temperature to crystallize 10% and 90% of the volume respectively. While the amorphous phase-change material provides good data retention (over 10 years) at room temperature, the crystallization can take place on a nanosecond time scale at elevated temperatures (purple curves). From [3] and [4].*

2.1 Growth of a crystallite

In the framework of classical crystallization theory the temperature dependence of the crystal growth velocity *u(T)* can be described as

$$u(T) \propto D(T) \cdot \left(1 - \exp\left(-\frac{\Delta G(T)}{k_B T}\right)\right) \tag{1}$$

The difference in Gibbs free energy between the liquid and crystal phase can be calculated using the Thompson-Spaepen approximation [11]

$$\Delta G(T) = \Delta H_m \frac{T_m - T}{T_m}\left(\frac{2T}{T_m + T}\right) \tag{2}$$

with heat of fusion ΔH_m and melting temperature T_m.

The moderate change of ΔG with temperature significantly below the melting temperature cannot explain the pronounced change in *u(T)* (see equation (2) and Fig. 3). Instead, the many orders of magnitude of change of crystallization speed (see both Equation (1) and Figure 3) originate mainly from the pronounced non-linearity in temperature dependence of atomic mobility *D(T)*, which is often expressed via the temperature dependence of viscosity *η(T)* using the anti-proportionality of the Stokes-Einstein equation:

$$D(T) \propto \frac{k_B T}{\eta(T)} \tag{3}$$

A breakdown of the Stokes-Einstein equation observed at temperatures close to T_g in various, especially fragile, supercooled liquids can be coped with a modification of the exponent $\xi \leq 1$ [10]:

$$D(T) \propto \frac{k_B T}{(\eta(T))^{\xi}} \tag{4}$$

This deviation has been commonly explained to originate from heterogeneous dynamics in supercooled liquids.

In recent years, evidence from different experimental methods pointed towards the existence of extraordinarily high fragilities in phase change materials [2,12]. The kinetic fragility is defined as the steepness of the viscosity in the supercooled liquid at the glass transition temperature T_g.

$$m = \left.\frac{\partial(\log_{10}\eta)}{\partial(T_g/T)}\right|_{T=T_g} \tag{5}$$

The according super-exponential deviation from a pure Arrhenius behavior in the temperature dependence of viscosity is traditionally described with the Vogel-Fulcher-Tamann equation:

$$\eta(T) = \eta_0 \cdot \exp\left(\frac{A}{T - T_0}\right) \tag{6}$$

with η_0, A and T_0 being constants.

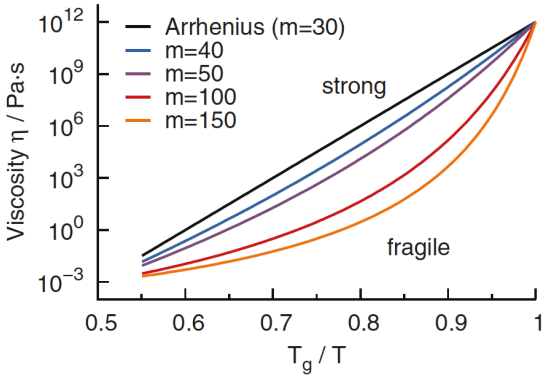


Fig. 4: *Illustration of the temperature dependence of the viscosity for varying fragility m according to the Vogel-Fulcher-Tammann equation (6) and the definition of fragility according to equation (5). From [3] and [4].*

With this high fragility of phase-change materials it is possible to explain the combination of a relatively wide window of fast crystallization at high temperatures with steeply increasing stability of the disordered phase towards lower/ambient temperatures, which is so beneficial for memory applications.

2.2 Nucleation

Without a preexisting crystal surface, the formation of a crystalline nucleus within the amorphous matrix is required, before the process of crystal growth can start. In classical nucleation theory, the nucleation rate of crystallites in a non-crystalline surrounding is derived beginning with the driving force for the formation of a spherical crystalline cluster of atoms, i.e. the difference in Gibbs free energy between the two phases:

$$\Delta G(r) = V(r) \cdot \Delta G_V + A(r)\sigma = \frac{4}{3}\pi r^3 \cdot \Delta G_V + 4\pi r^2 \sigma \tag{7}$$

Here, r is the radius, V the volume and A the surface of the nucleus, ΔG_V the difference in Gibbs free energy per unit volume and σ the interfacial energy per unit area. While the crystallization of a given volume leads to a reduction in total energy, the formation of a crystalline-to-amorphous interface costs energy. The critical radius r_C

$$r_C = \frac{2\sigma}{|\Delta G_V|} \tag{8}$$

describes the size of a crystalline cluster, at which the difference in Gibbs free energy (Equation 7) reaches its maximum. For nuclei larger than that, i.e. $r > r_C$, a further crystal growth is energetically favorable, whereas for $r < r_C$ the free energy of the system is increased by adding further atoms to the subcritical cluster. Nevertheless, the latter process takes place due to thermally induced statistical fluctuations, which eventually lead to the formation of supercritical crystalline nuclei. The steady-state nucleation rate I_{SS}, i.e. the number of supercritical nuclei forming per time and volume, is given by:

$$I_{SS} \propto \eta(T)^{-1} \exp\left(-\frac{\Delta G(r_C)}{k_B T}\right) = \eta(T)^{-1} \exp\left(-\frac{16\pi}{3k_B T}\frac{\sigma^3}{\Delta G_V} f(\theta)\right) \tag{9}$$

with $\eta(T)$ the temperature dependent viscosity. $f(\Theta)$ accounts for the effect of heterogeneous nucleation at interfaces with the wetting angle Θ. For homogeneous nucleation this factor equals unity.

2.3 Experimental quantification

In general, both crystal growth and nucleation can be simultaneously present during crystallization (see e.g. Fig. 5). Which of these processes is dominating the crystallization process, strongly depends on the observed volume and the presence of preexisting crystalline surfaces in the surrounding. Because of the fact that nucleation rate and growth velocity have differing temperature dependencies, e.g. maxima at different temperatures, the specific temperature distribution in a device can play an important role, too [13].

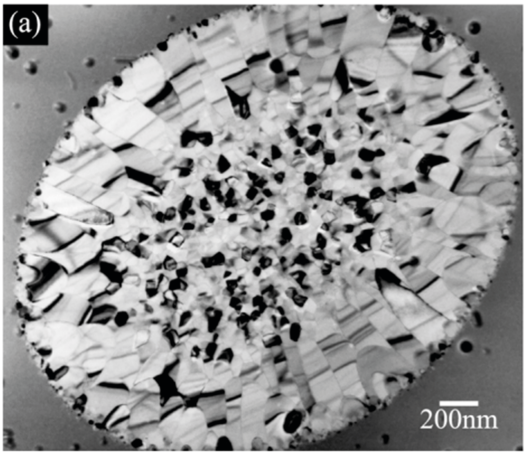


Fig. 5: *Transmission Electron Microscope image of a laser-crystallized spot in an otherwise as deposited-amorphous $Ge_2Sb_2Te_5$ film. While in the centre of the laser spot nucleation plays an important role, in the outer regions crystallization is dominated by crystal growth. From [3] and [7].*

In the past, measurements of crystal growth velocity have been limited to rather low temperatures where crystallization speeds are still slow. Until a few years ago fast measurements have always been performed in a non-isothermal way employing ultra-fast differential scanning calorimetry, short laser or voltage pulses to crystallize a small volume of material causing severe difficulties to obtain the temperature dependence of nucleation and growth velocities.

A new laser-based experimental approach allows the investigation of the technologically relevant melt-quenched amorphous phase under isothermal conditions covering a large range of crystal growth velocities (8 orders of magnitude) reaching up to the fastest regime (>1m/s). In this method time resolved reflectivity measurements with a bi-chromatic laser setup are employed. An intense laser pulse is used to locally melt a cylindrical volume with a radius of several hundred nm in a thin crystalline phase change film. This material is sandwiched between two layers of transparent ZnS-SiO_2 as depicted in the insets of figure 6. Once the laser pulse ends, the heat is very efficiently dissipated out of the melt into the Silicon substrate creating a melt-quenched mark. Finite element simulations of the layer stack show that it takes no longer than 100 ns to cool the phase change material down to the temperature of the substrate (blue curve in fig. 6). A second, low-intensity continuous-wave laser probes the reflectivity of the layer stack at exactly the same position where the first laser melts the phase change material (black curve in fig. 6). The reduction of reflectivity is a measure for how much of the previously crystalline material is amorphized. After thermalization, the gradual

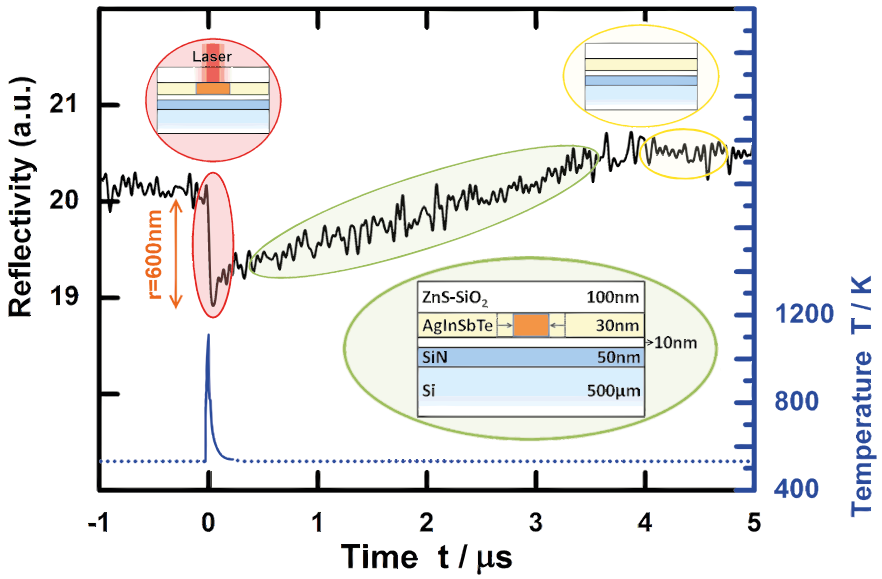


Fig. 6: *Time resolved reflectivity measurement and simulated temperature profile. The black line is the reflectivity trace collected during a recrystallization experiment performed at a substrate temperature of 533K. The zero of the timescale corresponds to the creation of the amorphous mark by the application of a laser pulse (83 mW for 30ns at 658nm wavelength). At this time the reflectivity suddenly decreases (red ellipse) and then, due to the recrystallization process (green ellipse), it increases again up to a steady state value (yellow ellipse) that corresponds to the complete recrystallization. The blue line shows the temperature profile during the laser irradiation process, simulated employing a finite element method. From [2].*

recovery back to a high reflectivity is thus an indicator for the progress of recrystallization. During the whole experiment, the sample is heated homogeneously at the temperature for which crystallization will be studied. The laser heating, however, is only used for the initialization of a crystallization experiment by creating an amorphous mark. This is crucial to obtain quantitative results for the temperature dependence of the crystal growth velocity using minimal assumptions.

The erasure of the amorphous bit in the film of phase change material takes place by growth from the rim as verified by TEM measurements. Thus the crystal growth velocity at a given temperature can be calculated by dividing the radius r of the created mark at the beginning of the recrystallization process by the time it takes for the reflectivity to fully recover. The initial radius r is computed from the drop in reflectivity induced by the laser pulse while taking into account the intensity profile of the probe laser and the dielectric function of all materials in the layer stack.

The investigated material shows an activation energy for crystallization of 2.7 eV within a temperature range between 418 K and 553 K (see Fig. 7). While at low temperatures (~420 K) the growth velocity was found to be in the range of 100 nm/s, it is strongly enhanced to more than 3 m/s at the upper temperature limit of the experimental range.

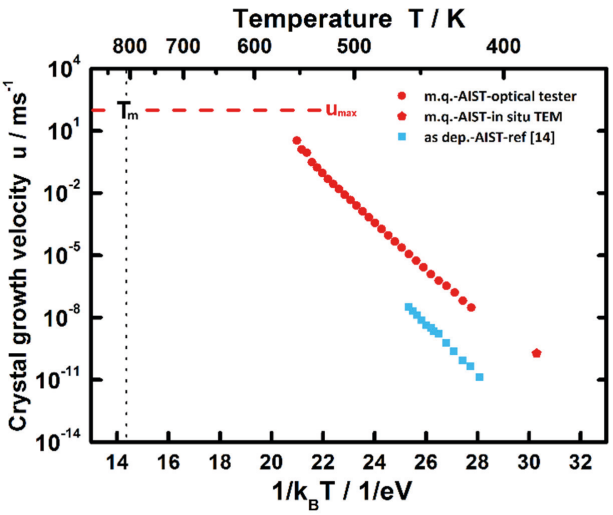


Fig. 7: *Temperature dependence of crystal growth rate. The growth velocity in melt-quenched amorphous AgInSbTe (red circles) as measured via time resolved reflectivity. The data exhibits an Arrhenius dependence on temperature characterized by a unique activation energy of 2.7 eV. A similar behaviour has been measured over a much smaller range of velocities and temperatures in as-deposited blanket AgInSbTe thin films [15] (blue squares). The maximum speed for melt-quenched AgInSbTe (~100 m/s) has been estimated on the basis of the two-pulses experiments. From [2].*

For a comprehensive interpretation of the crystallization kinetics one can derive the values for the viscosity η from the measured crystal growth velocities u employing the tight connection between both quantities (equations 1 and 3). The resulting temperature dependence of the viscosity is shown in Fig. 8. The values derived from reflectivity measurements represent a melt-quenched amorphous state for which the atomic configuration deviates from the equilibrium configuration of the supercooled liquid at a rather high temperature due to the vast cooling rates employed.

It is obvious from this diagram that the experimentally determined Arrhenius behaviour cannot extend to much higher temperatures, since the viscosity value at around 550 K (~170 mPas) is not even two orders of magnitude away from the viscosity measured in the liquid (~2 mPas). Such a pronounced flattening out of $\eta(T)$ (and thereby also of $u(T)$) towards higher temperatures can only be realized if a material's supercooled liquid phase has a high fragility. Fragilities reported in literature range from 20 for very strong liquids like SiO_2 up to over 150 for some very fragile, typically organic polymers (inorganic materials show fragilities up to m~90). As can be nicely seen in Fig. 8, in order to bring the viscosity of the supercooled liquid phase of AgInSbTe up from the viscosity of the liquid phase (around $10^{-3} Pa\ s$) in a way that it is in line with the values derived from the reflectivity measurements, extremely high fragilities of around 130 are necessary (lines in Fig. 8).

It is remarkable, that other materials with high fragility, i.e. typically organic/molecular compounds, generally show very slow crystallization kinetics because of their rather cumbersome building blocks that need to be rearranged, while the corresponding driving forces per unit (or per mol) are not so high (intermolecular interactions via van-der-Waals forces are rather

weak). So it seems that it is the unusual combination of low viscosity in the liquid, small building blocks, and significant driving forces (all not uncommon for inorganic materials), together with an extremely high fragility, that opens up a wide temperature window of low viscosity and thus, high crystallization speeds.

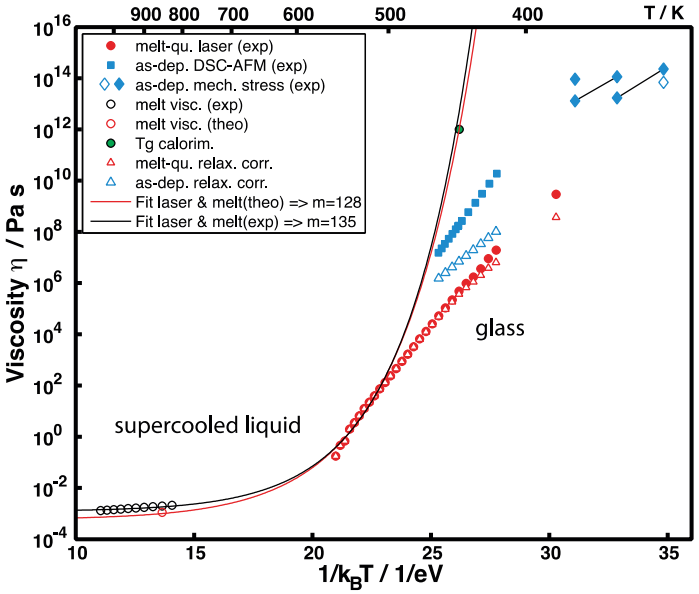


Fig. 8: *Temperature dependence of viscosity. Reversing equations (1)-(3) and using the growth velocity measurements of figure 7 the viscosity of AgInSbTe is calculated as a function of temperature (filled red circles). The general understanding that a glass is formed upon cooling from a supercooled liquid implies that the curves of supercooled liquid (continuous lines) and glass (red triangles) have to connect. The lines (black and red) are obtained by fitting the equation proposed by Mauro et al. [14] for the description of the viscosity of a supercooled liquid to the laser results at the 11 highest measured temperatures together with literature values for the viscosity in the liquid phase. Both fits correspond well with a viscosity of 10^{12} Pas at the glass transition temperature Tg = 443 K (green filled circle) that was previously observed for AgInSbTe using calorimetry [16]. The blue squares are obtained using the data on as-deposited AgInSbTe reported in ref [15]. The blue diamonds are extracted from ref. [17] in which viscosity of AgInSbTe has been measured via stress relaxation. The filled blue diamonds represent the viscosity of annealed amorphous samples, the open diamond states the original value before annealing. The original viscosity values derived from growth velocities (red filled circles from laser reflectivity experiments and blue filled squares from previous studies on as-deposited amorphous AgInSbTe) are corrected for structural relaxation they had time for during the experiments resulting in open red and open blue triangles, respectively. From [2].*

3 Threshold Switching

An increase of the temperature in a phase change memory element into the regime of fast crystallization needs sufficient Joule-heating and thus significant current densities. Since the resistivity of the amorphous phase is high and the available voltages in a memory device are typically limited to a few volts, the increase in local temperature could, in principal, pose a challenge to the ability to electrically switch a phase-change memory element. Thus, the existence of a pronounced non-linearity in the current-voltage characteristics of these amorphous materials is essential for a memory application. Especially the quite abrupt transition from a poorly to a highly conductive amorphous state at high electrical fields (threshold switching) is an enabling feature.

Since the discovery of the threshold-switching effect in 1968 by Ovshinsky [18] several different physical mechanisms have been proposed for its explanation. An extensive discussion about whether the driving force for this reversible switching phenomenon is controlled by a thermal or by an electronic effect was settled in the 1980s in favor of an electronic excitation mechanism. While today there is a broad agreement about the electrical-field-driven nature of the reversible threshold-switching, thermal effects are also considered, especially when dealing with nano-scale devices. In order to give an impression of the current discourse on the potential mechanisms behind threshold switching, three incompatible models are highlighted.

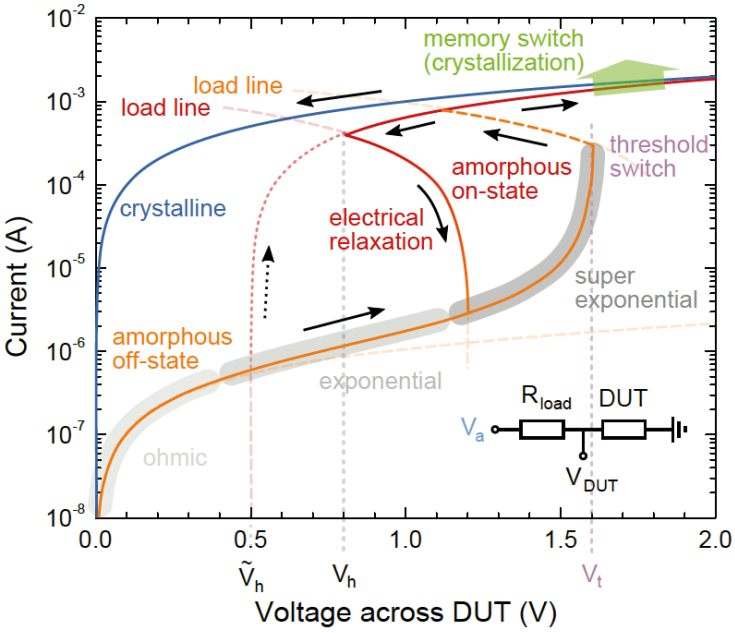


Fig. 9: *Illustration of a typical current-voltage-characteristic for phase change materials. In the logarithmic current scale the features of the amorphous off-state that are not visible in Fig. 2 can be seen. At low applied voltage the conductivity of the amorphous off-state is ohmic. At higher voltages the conductivity first increases exponentially and increases, at even higher voltages, super exponentially until the threshold switch occurs. From [3] and [5].*

Ielmini et al. proposed a widely cited model for the sub-threshold conduction and the transient effect of threshold switching in amorphous phase-change materials. [19] In their work, the non-linear current-voltage characteristic at moderate fields is explained by the Poole-Frenkel effect. Here, the potential energy barrier for excitation of trapped charge carriers into conducting states is reduced by the presence of near trap states and by the electrical field, which thereby strongly enhances the electrical conductivity. In 2007, this existing model for the steady-state transport in amorphous solids was extended by Ielmini et al. in order to also explain the regime at high electrical fields where threshold switching occurs. [19] Under the influence of strong fields charge carriers can be excited from deep trap states into shallow trap states close to the band edge. A relaxation mechanism counteracts the carrier generation leading towards a steady-state distribution. When enough carriers are excited into shallow trap states, threshold switching occurs.

The second model, which should be highlighted, was first proposed by Adler et al. in 1980 [20] and more recently reformulated by Pirovano et al. and Redaelli et al. [21,22]. Similar to Ielmini´s model, the threshold switching effect is described by an interaction of charge carrier generation and recombination mechanisms, whereas the electrical conduction is explained by trap-limited band transport. In case of low electrical fields, most of the generated free charge carriers can recombine via Shockley-Hall-Reed and/or Auger recombination. The generation mechanism (e.g. impact ionization and avalanche multiplication) strongly depends on the electrical field and the concentration of free carriers. In this model the breakdown in conductivity occurs at large electrical fields, when all traps close to the band edge are filled and the rate of recombination saturates. In this moment the process of recombination cannot counterbalance the generation of charge carriers anymore, leading to a strong increase in free charge carriers in the conduction band and thereby to threshold switching. Although both mechanisms for threshold switching proposed by Ielmini et al. and by Pirovano et al. appear to be similar on the first view, it has to be noted that the generation and recombination mechanisms as well as the detailed dependencies on the electrical field and charge carrier concentration differ significantly.

Thirdly, as an extension to the classical nucleation theory, Karpov et al. proposed a field-induced nucleation model to explain the threshold-switching phenomenon [23,24]. In this model, the electrostatic energy and thereby the free energy of a system is reduced by the formation of a crystalline nucleus within the amorphous material under bias. The energy barrier for the formation of a stable nucleus as well as the critical radius is reduced in the presence of an electrical field, which facilitates the nucleation process (see Figure 10). According to Karpov et al. the nucleation takes place heterogeneously at the interface between electrode and phase-change material. Once a crystalline nucleus is formed, it grows into a cylindrical filament. It thereby enhances the electrical field in the amorphous material at the tip of the filament accelerating further nucleation in this region. In the moment when the crystalline filament fully bridges the two electrodes, the threshold-switching becomes observable as a pronounced increase in conductance of the memory device. If the electrical field is removed before the crystallite has grown larger than the critical radius for the field-free case, the conductive filament decays. This way the field-induced reduction in the critical radius can explain the transient nature of the threshold switching event and the relaxation from the highly conductive amorphous “ON-state” back to the poorly conductive amorphous “OFF-state”.

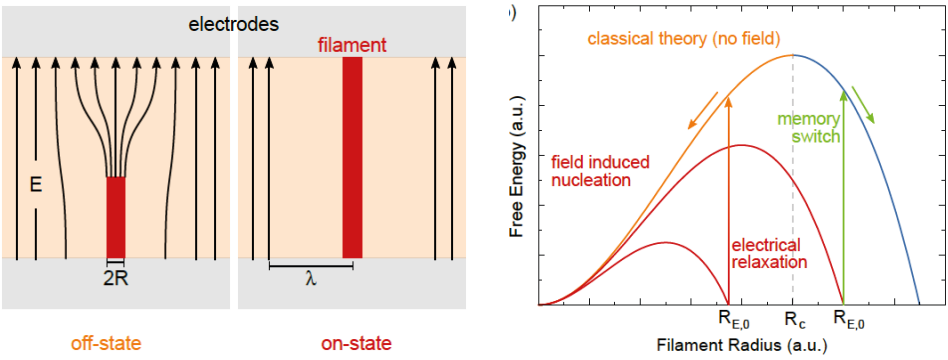


Fig. 10: *Illustration of the field-induced formation of a crystalline filament according to Karpov. Left: Once a nucleus is formed, the field is increased at the rim of the nucleus facilitating crystallization in field direction until a filament bridging the electrodes is formed. The free energy diagram (right) shows the field induced reduction of the critical radius R_c to form a nucleus. If in the on-state (red) the stable radius of the filament $R_{E,0}$ is smaller than the critical radius without a field R_c, the filament will disappear and the phase change material relaxes back into the off-state (orange). If the stable radius of the filament is bigger than the critical radius without field upon removal of the field, the crystalline filament is stable and can even grow further (blue). A memory switch has happened. From [5].*

4 Experimental observations in memory devices

The switching speed in phase-change memories is mainly limited by the SET process and the involved crystallization kinetics. In vertical phase change memory devices the amorphous plug is typically surrounded by crystalline phase-change material. Therefore, a crystalline-to-amorphous interface is always present in the highly resistive state. A study by Bruns et al. indicated that crystallization speed in phase-change materials such as GeTe is strongly dependent on the size of the amorphous region, which is represented by the device resistance (see Figure 11). With increasing device resistance (and plug size) the minimal pulse duration for full crystallization increases. This can be understood as experimental evidence for a growth-dominated crystallization process. Waiting for a spontaneous formation of supercritical nuclei is not necessary, which accelerates the total process of crystallization.

Besides the strong non-linearity in crystallization kinetics, the stability of the amorphous state also benefits from its non-linear current-voltage characteristics. Before any crystallization via Joule heating can take place, the reversible transition from the amorphous “OFF-” to the “ON-state”, the threshold switching, has to occur. Therefore, switching speed and retention of phase change memories are also determined by the threshold-switching phenomenon. Performing experiments on as-deposited amorphous lateral phase-change cells it has been shown that threshold switching happens, when the electrical field exceeds a critical value. [25] This critical field, however, is sensitive to the chosen pulse parameters and to the phase-change material under test.

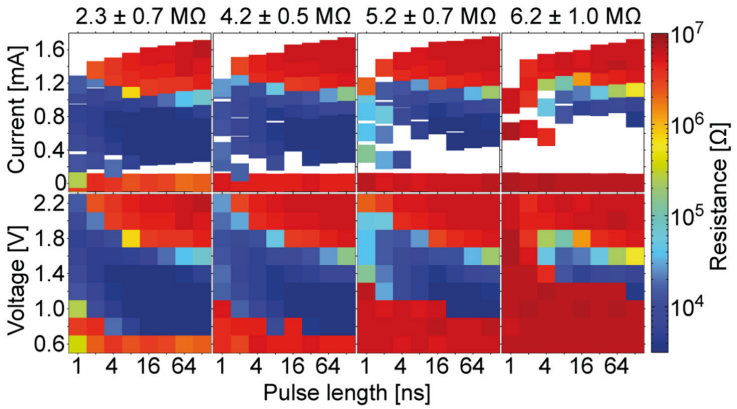


Fig. 11: *Crystallization speed of a vertical GeTe-mushroom cell programmed in four different initial high resistance states (columns). The color code shows how much the device resistance changed upon an excitation with a particular voltage pulse. In the bottom row, this resistance change is plotted dependent on the voltage and the duration of the applied set pulse. The top row additionally illustrates the dependence of this resistance change on the maximum current passing through the device during the pulsed electrical excitation.. Reamorphization takes place at currents larger than 1.1 mA. Crystallization can only occur when a minimal voltage (the threshold voltage) is exceeded (transition from red to blue data points). The comparison of experiments with four different initial RESET states reveals an increasing threshold voltage with device resistance (which is representative for the size of the amorphous region), indicating the field-driven nature of threshold switching. From [8].*

While delay times for threshold switching have been studied for almost 50 years now, new insights were reported very recently [9]. In that work, experimental evidence for an accumulative effect towards threshold switching was shown. Before the abrupt breakdown in resistivity, a continuous and linear increase in current ("pre-switching-slope", preSS) can be observed in case of sufficiently high electrical fields (Figure 12). For lower field strength no increase in current and also no switching event occurs. Thus, a significant preSS foretells the occurrence of the threshold switching effect. This way, a minimal electrical field for threshold switching can be experimentally determined for phase-change memories. The existence of such a minimal field ensures a high stability of the cell state during read-out and strengthens the data retention in amorphous phase-change materials.

While the threshold switching delay time was extensively studied as a function of applied voltage and cell resistance, the influence of ambient temperature is rarely characterized [24, 27, 28]. In the few existing works a general trend of switching at lower voltages was found for increasing temperatures.

In a recent work experiments with mushroom type memory cells were analysed with respect to the memory switching kinetics in doped $Ge_2Sb_2Te_5$ [26]. It was possible to derive an Arrhenius behaviour for the crystal growth velocity with an activation energy of 3 eV spanning over eight orders of magnitude. This study on nano-structured devices further strengthened the results from the laser reflectivity measurements mentioned above, which showed a pronounced non-linearity in crystallization kinetics ruled by an extraordinarily high fragility of the supercooled liquid.

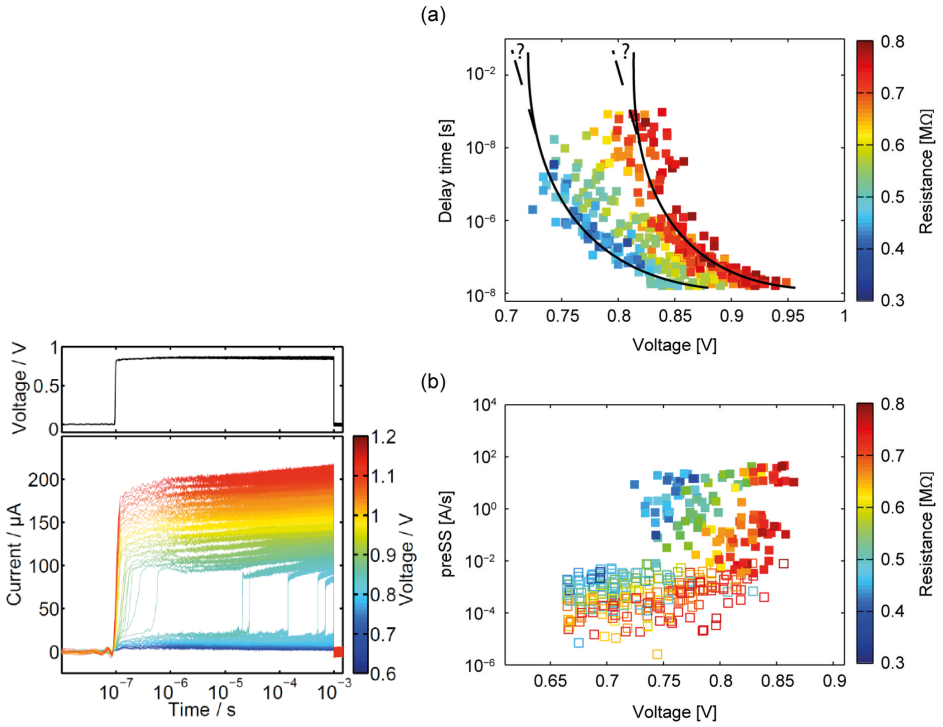


Fig. 12: ***Left:*** *Time-resolved voltage and current traces for various applied constant-voltage pulses on a logarithmic time scale. The current through the cell is depicted for multiple, stepwise increased constant-voltage pulses (amplitude represented by colour code). By increasing the voltage amplitude, a systematic shortening of the threshold switching delay time (which is indicated by a sharp increase in the current signal) can be observed.*
Right: *(a) Threshold switching delay time as a function of applied voltage for GeSbTe based devices with varying initial amorphous plug size (i.e. initial resistance between 300 and 800 kΩ). In agreement with an electric-field-induced switching-mechanism a decrease in initial cell resistance, and thus a decrease in amorphous thickness between effective electrodes, results in a lowering of the voltages, at which threshold switching takes place after a specific delay time. Although delay times were measured over five orders of magnitude in time, the existence of a minimal threshold field cannot be proven based on such an analysis alone.*
(b) Pre-switching-slope as a function of applied voltage measured for various initial cell resistances. Filled squares indicate the existence of a threshold-switching event during the applied pulse, whereas unfilled symbols represent experiments, in which no switching event occurred. The device resistance shows a pronounced influence on the voltage-dependent pre-switching-slope. For smaller initial resistances the curves are shifted towards lower voltages. Because a smaller initial resistance is indicative of a smaller size of the initial amorphous plug, this shift of curves is in line with the increase in current prior to threshold-switching being an effect induced by the electric field. From [9].

Revisiting the mechanisms described above, the combination of long-term stability and ultra-fast switching in memories based on phase change materials is cooperatively achieved by the non-linear current-voltage characteristics including the effect of threshold switching as well as by the super-exponential crystallization kinetics. At voltages well above the minimal electrical field for switching, the threshold switching event can be triggered within less than a nanosecond and the extremely large crystal growth velocities (> 1 m/s) enable an ultra-fast SET process for this class of material. At low voltages, below the minimal threshold field, good data retention is guaranteed over long time scales due to the high activation energies that need to be overcome for crystallization.

References

These lecture notes were compiled from the following publications:

[1] S. Menzel, U. Böttger, M. Wimmer and M. Salinga (2015), *Physics of the Switching Kinetics in Resistive Memories*. Advanced Functional Materials, 25: 6306–6325.

[2] M. Salinga, E. Carria, A. Kaldenbach, M. Bornhoefft, J. Benke, J. Mayer, M. Wuttig, Nature Communications 2013, *4*, 2371.

[3] D. Lencer, M. Salinga, M. Wuttig, *Design Rules for Phase-Change Materials in Data Storage Applications*, Advanced Materials 2011, *23*, 2030.

[4] D. Lencer*, Design Rules, Local Structure and Lattice-Dynamics of Phase-Change Materials for Data Storage Applications,* Doctoral thesis, RWTH Aachen, 2010.

[5] D. Krebs, *Electrical Transport and Switching in Phase Change Materials,* Doctoral thesis, RWTH Aachen, 2010.

[6] M. Salinga, *Phase Change Materials for Non-volatile Electronic Memories,* Doctoral thesis, RWTH Aachen, 2008.

Other references:

[7] I. Friedrich, V. Weidenhof, S. Lenk, M. Wuttig, Thin Solid Films 2001, 389, 239.

[8] G. Bruns, P. Merkelbach, C. Schlockermann, M. Salinga, M. Wuttig, T. D. Happ, J. B. Philipp, M. Kund, Applied Physics Letters. 2009, 95, 043108.

[9] M. Wimmer, M. Salinga, The gradual nature of threshold switching. New Journal of Physics 2014, 16, 113044.

[10] M. D. Ediger, Annu. Rev. Phys. Chem. 2000, 51, 99.

[11] C. V. Thompson, F. Spaepen, Acta Metallurgica 1979, 27, 1855.

[12] J. Orava, A. L. Greer, B. Gholipour, D. W. Hewak, C. E. Smith, Nat. Mater. 2012, 11, 279.

[13] G. W. Burr, P. Tchoulfian, T. Topuria, C. Nyffeler, K. Virwani, A. Padilla, R. M. Shelby, M. Eskandari, B. Jackson, B.-S. Lee, J. Appl. Phys. 2012, 111, 104308/1.

[14] Mauro, J.C., Yue, Y.Z., Ellison, A.J., Gupta, P.K. & Allan, D.C. Viscosity of glass-forming liquids. P Natl Acad Sci USA 106, 19780-19784 (2009).

[15] J. Kalb, F. Spaepen, M. Wuttig. Appl Phys Lett 84, 5240-5242 (2004).

[16] J. Kalb, M. Wuttig, F. Spaepen. J Mater Res 22, 748-754 (2007).

[17] J. Kalb, F. Spaepen, T.P.L. Pedersen, M. Wuttig. J Appl Phys 94, 4908-4912 (2003).

[18] S. R. Ovshinsky, Phys. Rev. Lett. 1968, 21, 1450.

[19] D. Ielmini, Y. Zhang, J. Appl. Phys. 2007, 102, 054517.

[20] D. Adler, M. S. Shur, M. Silver, S. R. Ovshinsky, J. Appl. Phys. 1980, 51, 3289.

[21] A. Pirovano, A. L. Lacaita, A. Benvenuti, F. Pellizzer, R. Bez, IEEE Transactions on Electron Devices, USA 2004, 51, 452.

[22] A. Redaelli, A. Pirovano, A. Benvenuti, A. L. Lacaita, J. Appl. Phys. 2008, 103, 111101/1.

[23] V. G. Karpov, Y. A. Kryukov, I. V. Karpov, M. Mitra, Phys. Rev. B: Condens. Matter 2008, 78, 52201/1.

[24] V. G. Karpov, Y. A. Kryukov, M. Mitra, I. V. Karpov, J. Appl. Phys. 2008, 104, 054507.

[25] D. Krebs, S. Raoux, C. T. Rettner, G. W. Burr, M. Salinga, M. Wuttig, Appl. Phys. Lett. 2009, 95, 82101/1.

[26] A. Sebastian, M. Le Gallo, D. Krebs, Nature Communications 2014, 5, 4314/1.

[27] S. H. Lee, H. K. Henisch, J. Non-Cryst. Solids 1972, 11, 192.

[28] S. Lee, D. S. Jeong, J.-H. Jeong, W. Zhe, Y.-W. Park, H.-W. Ahn, B.-.K. Cheong, Appl. Phys. Lett. 2010, 96, 23501/1.

D 8 Interfacial Phase Change Materials

Riccardo Mazzarello
Institute for Theoretical Solid State Physics
RWTH Aachen, 52074 Aachen

Contents

1 Introduction **2**

2 First experimental evidences **3**

3 Properties of interfacial phase-change materials **5**
3.1 Structure 5
3.2 Switching mechanism 7
3.3 Electronic properties 9

4 Latest developments **11**

Lecture Notes of the 47^{th} IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

Phase-change materials (PCMs) possess an outstanding combination of properties [1]: their amorphous and crystalline state exhibit strong optical and electronic contrast, furthermore the amorphous phase is stable for decades at room temperature but crystallizes extremely fast at moderately high temperatures (typically, 600-700 K). These properties are currently exploited in rewritable optical devices such as rewritable Blu-Ray discs and in non-volatile phase-change memories. In these devices, crystallization of a phase-change cell is induced by applying locally intermediate-power laser or current pulses, so as to increase temperature above the crystallization temperature. The transition from the crystalline to the amorphous state is instead obtained by applying a short, intense pulse to melt the system, followed by rapid quenching from the melt.

Phase-change memories are a very promising technology, nevertheless their performance has to be further improved to win the market. Although fast, crystallization is still the slower process which affects the maximum speed of the device. On the other hand, amorphization is the more energy-consuming process, because relatively large currents are needed to melt the system. Recently, a novel class of PCMs has emerged [2, 3, 4], which holds the promise of overcoming these two drawbacks. Reversible transitions between two states with different resistivities occur in these materials as well. However, switching between the two phases turns out to be faster and energetically less demanding than the amorphous/crystalline transitions in standard PCMs. It has been suggested that these properties stem from the fact that a) the two relevant states are both crystalline and b) the transitions are constrained to atomic motion in one dimension [4]. This new family of PCMs has been called interfacial PCMs (IPCMs).

Interestingly, IPCMs consist of the same atomic elements (and even the same building blocks) of the technologically most important family of standard PCMs, namely the GeSbTe (GST) compounds which lie along the GeTe-Sb_2Te_3 pseudobinary line [5]. More specifically, IPCMs have been obtained by building superlattice structures made of GeTe and Sb_2Te_3 layers [4].

In spite of successful experimental demonstrations of memory devices based on the new switching mechanism, this very mechanism has not been fully elucidated. In fact, the triggering of the structural transitions has been ascribed to several different effects, including electric field-induced effects [6, 7] and thermal activation [8]. Even the atomic structure of the two relevant states, the low-resistance SET state and the high-resistance RESET state, is still under debate. Accurate experimental determination of the layer sequence has proven to be challenging, owing to comparable interlayer distances in different structures, the thinness of the building blocks, and the predominant use of sputtering techniques to grow the samples, which typically results in a high density of structural defects.

It has also been suggested that one of the states involved in the transition exhibits non-trivial topological properties [6] and, thus, the structural switching may also bring about a transition between two topologically distinct states. This fact links IPCMs with topological insulators [9] and semimetals [10], an extremely active field at the forefront of condensed mattery physics.

In the following two sections, we provide an introduction to important experimental findings about IPCMs and to theoretical work aiming at elucidating their structural, electronic and kinetic properties. Finally, in the last section, we discuss very recent developments, which partially challenge our understanding of these materials and show that this fascinating field of research is still in its infancy.

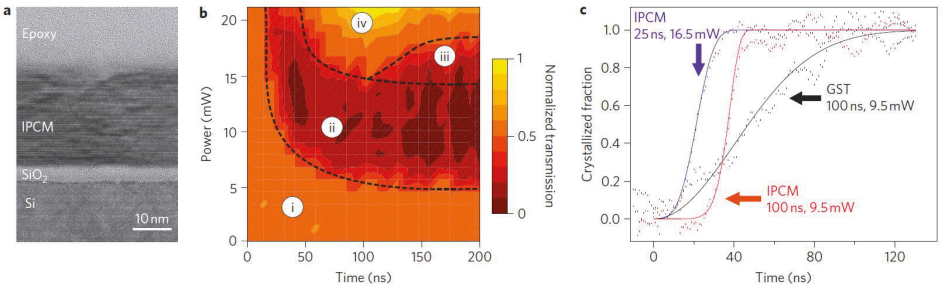


Fig. 1: *(a) High-resolution transmission electron micrograph image of an as-grown $(GeTe)_2(Sb_2Te_3)_4$ IPCM on silicon. Assuming that no intermixing between GeTe and Sb_2Te_3 occurs, the thickness of the $(GeTe)_2$ layer and the $(Sb_2Te_3)_4$ layer is 1 nm and 4 nm, respectively. (b) Time-resolved pump-probe static tester measurement for the transition from the RESET to SET state of a mark of radius 400 nm in the IPCM film. The RESET mark was created with a 40 ns laser pulse with a power of 32 mW. The figure shows the normalized optical transmission of a 100 μW probe beam through the RESET mark as a function of time, during and after the 100 ns laser pump pulse, for different incident optical powers. Regions (i-iv) indicate, respectively: no change in transmission (i), a reduction in transmission due to the RESET to SET transition (ii), an increase followed by a reduction in transmission, indicative of melting and subsequent crystallization (iii), and an increase in transmission, due to melting and partial ablation (iv). (c) Re-crystallized fraction of a RESET mark (created with a 40 ns laser pulse with a power of 32 mW) as a function of time for GST using 100 ns pump pulses with power 9.5 mW (black line) and for IPCM using 100 ns pump pulses with power 9.5 mW (red line) and 25 ns pulses with power 16.5 mW (blue line). Reprinted by permission from Macmillan Publishers Ltd: Ref. [4].*

2 First experimental evidences

The fabrication of GeTe-Sb_2Te_3 superlattices was first reported in Ref. [2]. In this work, it was shown that, by applying appropriate electrical or laser pulses, it is possible to induce reversible transitions between three different states, namely an amorphous superlattice, a crystalline one, and a mixed superlattice consisting of alternating amorphous (GeTe) and crystalline (Sb_2Te_3) component materials. More specifically, the transition from the amorphous to the intermediate mixed state could be obtained by applying a less intense pulse than that needed to fully crystallize the system. This behaviour stems from the fact that Sb_2Te_3 has a lower crystallization temperature than GeTe.

The superiority of GeTe-Sb_2Te_3 superlattices in terms of switching speed and energy consumption as compared to standard PCMs was first demonstrated in the ground-breaking work by Simpson *et al.* [4]. The authors used a physical vapour deposition system to grow GeTe-Sb_2Te_3 superlattices on a silicon substrate (see Figure 1(a)). The thickness of the GeTe and Sb_2Te_3 layers ranged between 5 Å and 40 Å. According to the authors, no evidence for intermixing of GeTe and Sb_2Te_3 was found. The switching time for IPCM samples and standard GST films was measured optically using a laser pump-probe static tester system. The change in optical transmission through a RESET mark in the IPCM film as a function of time, during and after the application of a 100 ns laser pump pulse with varying incident optical powers, is shown in Figure 1(b). Figure 1(c) demonstrates that, for low-power (9.5 mW) laser pulses of 100 ns

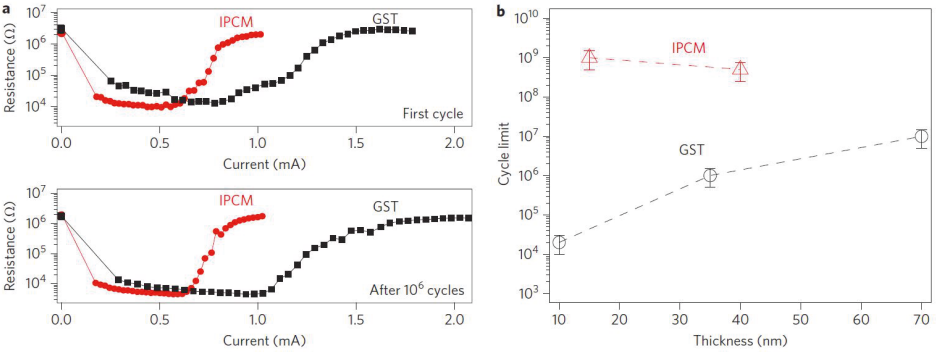


Fig. 2: *(a) Electrical switching characteristics of phase-change memory cells. (a) Plots of resistance versus current for GST (black squares) and IPCM (red circles) devices in the first cycle (upper panel) and after 10^6 cycles (lower panel). The IPCM employed in the second device was $(GeTe)_4(Sb_2Te_3)_2$. The pulse length for the SET operation was 50 ns and 100 ns for the IPCM and GST device, respectively. The RESET pulse length was set to 50 ns for both devices. (b) Plot of the maximum number of SET-RESET cycles as a function of film thickness. IPCM cells exhibit better cyclability than cells based on GST, as well as less pronounced dependence on thickness. Reprinted by permission from Macmillan Publishers Ltd: Ref. [4].*

duration, the transition rate to the SET state of the ICPM sample was four times higher than that of the GST film. Furthermore, for higher-power (16.5 mW) laser pulses of 25 ns duration, the onset of crystallization of the IPCM was extremely fast and complete crystallization occurred without any subsequent damage.

Simpson *et al*. [4] also investigated prototypes of electrical solid-state memory cells containing IPCMs. They showed that the currents required to reversibly switch IPCM devices between the SET and RESET states are lower than those needed for standard GST-based devices; see Figure 2(a). In particular, the energy to be provided to trigger the transition to the SET state is much smaller for IPCM memory cells (11 pJ versus 90 pJ). Interestingly, IPCM-based cells also displayed a more abrupt switch between the SET and RESET state, resulting in a less broad distribution of device characteristics. Moreover, the IPCM devices showed 1-2 orders of magnitude higher cyclability as compared to standard GST devices (10^8-10^9 versus 10^7), as shown in Figure 2(b).

Simpson *et al*. [4] measured the thermal conductivity of both ICPM and GST films and found that the conductivity of GST is lower than that of ICPM, which rules out reduction of thermal conductivity as the cause for the better performance of IPCM devices. They instead argued that the superior switching properties of IPCM memory cells stem from the fact that the RESET state of IPCM is crystalline and has a lower entropy than the amorphous RESET state of GST. Transmission electron microscopy images of the IPCM cell directly above the heating electrode after transition to the RESET state indicated that the layered structure was preserved and suggested that no amorphization process occurred. On the other hand, the reset of the IPCM cell with the same high-power electrical pulses used to reset GST led to a melt-amorphized region above the electrode [4].

Finally, it was conjectured in this work that the transition between the SET and RESET state in IPCM involves the motion of Ge layers at the Sb_2Te_3 interface; hence the name “interfacial

phase-change materials".
These important findings triggered intense experimental [11, 12, 13, 14] and theoretical [6, 15] efforts to elucidate the structural and electronic properties of the SET and RESET state and the kinetics associated with the switching mechanism, as discussed in the next section.

3 Properties of interfacial phase-change materials

3.1 Structure

As mentioned in the previous section, the transmission electron microscopy experiments performed in Ref. [4] suggest that both SET and RESET states are crystalline. This hypothesis is compatible with coherent phonon spectroscopy measurements carried out by Makino *et al.* [13]. In this work, both IPCM (namely, $(GeTe)_2(Sb_2Te_3)_4$ and $(GeTe)_2(Sb_2Te_3)$ superlattices) and conventional $Ge_2Sb_2Te_5$ films were considered. At low temperature (25 °C), coherent phonon signals were observed in the RESET state of both sets of samples. However, if temperature was increased to 180 °C, coherent phonon signals were significantly suppressed for the case of $Ge_2Sb_2Te_5$ films, owing to the transition from the amorphous to the crystalline state. On the contrary, coherent signals were still detected in the IPCM films, in spite of the induced RESET-to-SET phase transition. Furthermore, upon subsequent cooling of the samples, the attenuation of coherent phonons was found to be irreversible in $Ge_2Sb_2Te_5$ alloys (due to the irreversibility of the amorphous-crystalline transition), whereas, for IPCM, the intensity of the coherent phonon oscillations mostly recovered [13]. The authors attributed this behaviour to the fact that the RESET-SET transition of IPCMs involves two crystalline states and, thus, smaller atomic rearrangements as compared to conventional $Ge_2Sb_2Te_5$.
The experimental findings discussed so far call for an atomistic understanding of the structure of the relevant phases. To achieve this goal, it is crucial to complement experimental data with simulations. In particular, density functional theory (DFT) simulations [17, 18] are the ideal tool to compute energy differences between different structures and determine the most stable ones. Indeed, there have recently been numerous DFT studies of IPCM, which have mainly focused on $(GeTe)_2(Sb_2Te_3)$ superstructures. In the following, we also restrict ourselves to the latter systems.
Four ordered configurations of $(GeTe)_2(Sb_2Te_3)$ have been thoroughly investigated by DFT methods: these configurations have been called Kooi phase [19], Petrov phase [20], inverted Petrov phase [6] and ferroelectric phase [6], respectively. The four configurations are shown in Figure 3. Their structural properties can be better understood by first considering the two parent compounds, GeTe and Sb_2Te_3.
Bulk Sb_2Te_3 has a rhombohedral geometry but its structure can be visualized more easily in the conventional hexagonal supercell. It consists of quintuple (Te-Sb-Te-Sb-Te) layers stacked along the c direction of the supercell. The coupling between quintuple layers is of van der Waals type; for this reason, the gap between Te-Te layers is called "van der Waals gap". Crystalline GeTe has a rhombohedral geometry as well, which originates from a Peierls-like distortion of a rocksalt structure. The peculiar bonding in this structure has been described as resonant (although it is not perfectly resonant owing to said Peierls distortion) [21, 22]. GeTe is ferroelectric, in that the distortion induces an intrinsic dipole moment oriented along the $\langle 111 \rangle$ direction of the crystal.
Moving back to $(GeTe)_2(Sb_2Te_3)$, the ferroelectric phase consists of alternating Sb_2Te_3 quin-

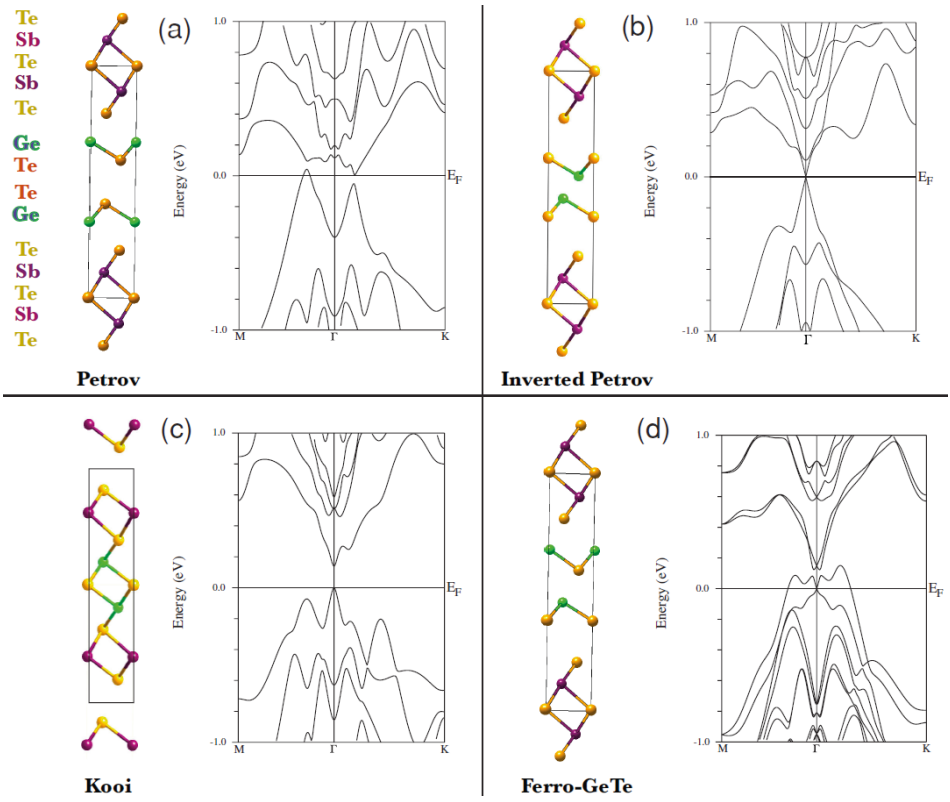


Fig. 3: *Proposed configurations for the IPCM $(GeTe)_2$-Sb_2Te_3 and corresponding bulk band structures along the M-Γ-K direction of the Brillouin zone, as computed in Ref. [6]. Green, dark magenta and orange spheres denote Ge, Sb and Te atoms, respectively. Notice that the inverted Petrov sequence displays two bulk Dirac-like cones at the Γ point, which touch at the Fermi energy E_F, leading to semimetallic behaviour. Nevertheless, this semimetallic phase is not protected by topology [10]. As a consequence, perturbations like strain and pressure can eliminate the degeneracy at the Dirac point and induce a gap. Reprinted from Ref. [6].*

tuple layers and double GeTe layers with the same orientation of the dipole moments as in the bulk GeTe phase. In the Petrov sequence, the $(GeTe)_2$ block has instead an antiferroelectric Ge-Te-Te-Ge sequence and the van der Waals gap lies between the Te layers in this block. In the inverted Petrov sequence, $(GeTe)_2$ forms the antiferroelectric sequence Te-Ge-Ge-Te. In this model, the van der Waals gap is between the $(GeTe)_2$ and (Sb_2Te_3) blocks. The Kooi phase is obtained by incorporating GeTe in the Sb_2Te_3 quintuple layer, resulting in the periodically repeated sequence Te-Sb-Te-Ge-Te-Ge-Te-Sb-Te. There is a van der Waals gap between Te-Te layers, analogously to the case of bulk Sb_2Te_3.
At zero temperature, the Kooi phase is the energetically most favourable configuration [6, 15, 23]. The Petrov and inverted Petrov sequence have intermediate (quasi-degenerate) energies, whereas the ferroelectric phase is the least stable one [6, 15] (in Ref. [14], it is instead reported that the ferroelectric sequence is the most stable among the latter 3 structures. Since few computational details are provided in this work, it is difficult to determine what the discrepancy is due to). However, the energy differences between these configurations are small, of the order of 0.1-0.2 eV per cell [6, 15]. Furthermore, the energy differences between the stable clean phase and the corresponding disordered phases characterized by compositional Ge-Sb disorder (i.e. intermixing of Ge and Sb layers) are also tiny: in fact, at room temperature, they are comparable to the configurational entropy contribution of the disordered phases [24] (similar considerations hold for other stoichiometries, such as $GeSb_2Te_4$ [25]). This suggests that Ge-Sb layer intermixing may occur in these systems.
Interestingly, the relative stability between the ordered structures turns out to depend on temperature, as discussed in Ref. [6, 15]. Nevertheless, there are some apparent discrepancies between these two works with respect to the stable high-temperature phases. More specifically, Ref. [6] predicted that, at 500 K, the ferroelectric and inverted Petrov phase have the lowest (average) energy. On the other hand, Ref. [15] presented calculations of the enthalpy as a function of temperature, based on the *ab initio* evaluation of the phonon dispersion spectrum, and showed that, above T = 125 K, the ferroelectic phase indeed becomes the most favorable one, however the inverted Petrov has higher entalpy, comparable to that of the Petrov and Kooi phase [15] (see Figure 4).

3.2 Switching mechanism

Recently, two models have been proposed to explain the switching mechanism. The first model describes the switching as due to reversible transitions between the ferroelectric (low-resistance SET state) and the inverted Petrov structure (high-resistance RESET state) [6, 8, 13]. This model was developed from the analysis of experimental data about the two states, including electron microscope images and diffraction experiments, as well as atomistic simulations. Ohyanagi *et al.* instead proposed a transition between the Petrov phase (SET state) and the inverted Petrov stacking (RESET state) [14]. In the latter work, IPCM films prepared by sputtering at different deposition temperatures were investigated. X-ray diffraction (XRD) measurements indicated that the films deposited at low temperature (200 °C) formed the ferroelectric phase. These samples showed no resistance change during SET and RESET operations. Based on first-principles simulations at T = 0, the authors attributed this behaviour to the high energetic stability of the ferroelectric state (on the contrary, the simulations of Refs. [6, 15] indicate that the ferroelectric phase is the least stable at low temperature, but has the lowest enthalpy at higher temperature above T = 125 K [15]: see relevant discussion in the previous subsection). On the other hand, films deposited at higher temperature (240 °C) displayed phase-change be-

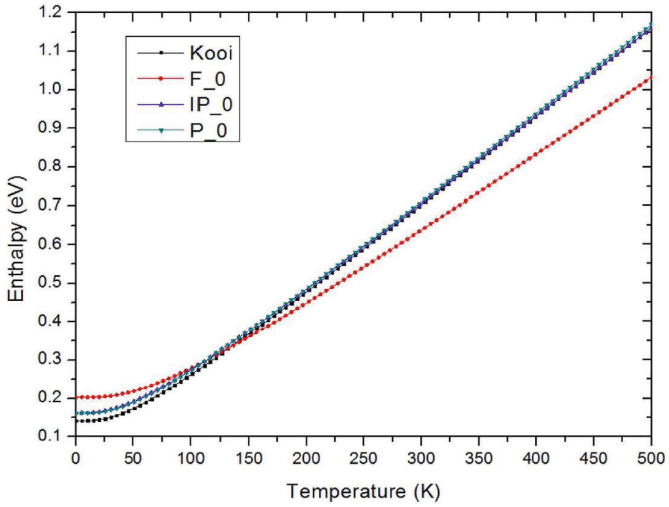


Fig. 4: *Plot of the enthalpy as a function of temperature for the 4 structures shown in Fig. 3. F_0, P_0 and IP_0 denote the ferroelectric, Petrov and inverted Petrov phase, respectively. Although, at low temperature, the Kooi structure has the lowest enthalpy, at temperatures above 125 K the ferroelectric phase becomes stable. Reprinted from Ref. [15].*

haviour and pronounced resistivity contrast [14]. Analysis of XRD measurements and *ab initio* simulations led the authors to the conclusion that, in these films, the switching occurred between the metastable Petrov and inverted Petrov structure. However, it remains unclear why the ferroelectric phase should not form at higher temperatures, if it is more stable than the other structures.

In both models, switching involves a vertical displacement of Ge layers through a Te layer, as well as an additional lateral movement of Ge atoms (this can be understood by considering that all the models exhibit *abc*-type stacking of the atomic layers). More precisely, model 1 and model 2 involve single and double flipping of Ge layers, respectively. In Ref. [15], the microscopic displacements of the Ge atoms were investigated by DFT techniques and the energy barriers for migration were computed for both models using the transition state search algorithm [16]. Since, as already mentioned, a vertical movement of Ge atoms from the initial low-resistance (ferroelectric or Petrov) state does not yield the inverted Petrov state (and vice versa), the authors of this work first classified the structures one can obtain from the four basic structures by changing the intra-layer orderings and calculated their energy. They found that all of the new structures have higher energies than the corresponding original structures (energy differences are typically of the order of tenths of eV per cell). Subsequently, Yu and Robertson calculated the energy barriers for the vertical displacement [15]. During this process, the Ge and Te layers cross each other. Not surprisingly, the maximum energy along the transition path corresponds to a configuration where the Ge and Te atoms lie in the same plane and the distance between nearest neighbour atoms is lowest. For model 1, the energy barriers are 2.84 eV (SET operation) and 2.56 eV (RESET), whereas, for model 2, they amount to 3.10 eV (SET operation) and 2.59 eV (RESET), respectively.

As far as the lateral movement of the Ge atoms is concerned, it turns out that there are two

possible ways for nearest neighbour Ge and Te atoms in a GeTe layer to exchange their positions so as to recover the abc stacking. In the first process, an atom moves over the top of an adjacent atom, resulting in a compression of the the GeTe and Sb_2Te_3 layers and the breaking of 2 out of 3 bonds with its neighbours. In the second case, the atoms instead move in-plane, breaking only one bond [15]. For this reason, the second process is characterized by lower energy barriers, ranging between 0.05 eV and 0.62 eV (the energy barriers of the first process are about 0.5 eV higher).
In summary, Ref. [15] showed that a) the switching between the crystalline states proposed in model 1 [6, 8, 13] and model 2 [14] is a two-step process, and b) the first process, i.e. the vertical flip of the Ge and Te atoms, has a higher energy barrier ranging between 2.56 and 3.10 eV.
In the next subsection, we discuss the electronic properties of the 3 crystalline states relevant to the 2 models, as well as of the Kooi structure.

3.3 Electronic properties

GeTe-Sb_2Te_3 superlattices have remarkable electronic properties. The building block Sb_2Te_3 is known to be a 3-dimensional [26] topological insulator [27]. Topological insulators are a recently discovered state of matter, characterized by a bulk band gap and conducting surface states [9]. The surface states exhibit spin-momentum locking with spin direction perpendicular to momentum and their gaplessness is topologically protected against perturbations which do not break time-reversal symmetry, such as non-magnetic disorder. These properties originate from the interplay between strong spin-orbit coupling and time-reversal symmetry. Assuming rotational invariance and a surface perpendicular to z, the surface states can be described by the Dirac-like Hamiltonian [9]

$$H_{surf} = v_F(\sigma^x k_y - \sigma^y k_x), \tag{1}$$

where v_F is the Fermi velocity and σ^x and σ^y are Pauli matrices acting onto the spin (we set $\hbar = 1$). Far from the Dirac point, deviations from the linear behaviour can occur.
Earlier studies of the topological properties of $(GeTe)_2$-Sb_2Te_3 predicted that the Petrov structure is a topological insulator, whereas the Kooi phase is a conventional band insulator [28]. Later, the same authors investigated topological phase transitions in ternary chalcogen GeSbTe and GeBiTe compounds [29]. They showed that, for GeSbTe materials with Kooi-like stacking order, a transition from normal insulator (observed in GeTe-rich materials) to topological insulator (found in Sb_2Te_3-rich compounds) occurs as a function of stoichiometry. The transition point was estimated to be between $GeSb_2Te_4$ and $Ge_2Sb_4Te_8$. Furthermore, they rationalized their results in terms of superlattice models of topological and band insulator layers. By employing model Hamiltonians, they showed that the electronic properties of these systems depend on the helicity ordering formed by the Dirac fermions in the superlattice. The helicity operator $\hat{h}$ is defined as:

$$\hat{h} = \frac{1}{k}(\sigma^x k_y - \sigma^y k_x). \tag{2}$$

The eigenvalues of $\hat{h}$ are ± 1, which, for a given eigenstate, correspond to momentum $\mathbf{k}$ parallel or antiparallel to $\hat{z} \times \boldsymbol{\sigma}$, respectively. For a given surface, the states belonging to the upper (or lower) Dirac cone have the same helicity but the two cones have opposite helicity. Positive (respectively negative) helicity corresponds to a cone having clockwise (resp. counterclockwise) chirality. The upper (lower) cones on the top and bottom surface of a topological insulator slab

have opposite helicities. The helicity of an isolated topological insulator is fixed; nevertheless, different types of helicity can occur inside a superlattice, in the presence of a finite coupling between surface Dirac fermions of different layers. The DFT simulations by Kim *et al.* [29] explicitly demonstrated that two different types of helicity ordering occur in the standard insulator $GeSb_2Te_4$ (counterhelicity ordering) and the topological insulator $Ge_2Sb_4Te_8$ (cohelicity ordering).

More recently, a comprehensive analysis of the electronic structure of the 4 stacking sequences of $(GeTe)_2$-Sb_2Te_3 was carried out in Ref. [6]. In this work, it was reported that, for inverted Petrov stacking order, $(GeTe)_2$-Sb_2Te_3 is a Dirac semimetal possessing 3-dimensional Dirac cones which touch at the Γ point of the Brillouin zone (see Fig. 3). The Dirac point is located exactly at the Fermi energy. To shed light on this behaviour, they also considered superlattice models consisting of alternating topological and band insulator layers, following a recent work by Burkov and Balents [10]. In the latter work, an effective Hamiltonian for such superlattices was defined, which contains a) standard Dirac-like terms describing the surface states of the topologically non-trivial layers and b) additional tunneling terms between neighbouring surface states. Assuming that the growth direction of the superlattice is parallel to z, the Hamiltonian reads:

$$H_{SL} = \sum_{\mathbf{k}_\perp} \sum_{ij} \left[v_F \tau^z (\sigma^x k_y - \sigma^y k_x) \delta_{i,j} + \Delta_S \tau^x \delta_{i,j} + \frac{1}{2} \Delta_D (\tau^+ \delta_{i,j+1} + \tau^- \delta_{i,j-1}) \right] c^\dagger_{\mathbf{k}_\perp,i} c_{\mathbf{k}_\perp,j}, \tag{3}$$

where the indices i and j label the topological insulator layers, τ are the Pauli matrices which act onto the surface index (top/bottom), $c^\dagger_{\mathbf{k}_\perp,i}$ ($c_{\mathbf{k}_\perp,i}$) is the creation (annihilation) operator of electrons with in-plane wavevector $\mathbf{k}_\perp \equiv (k_x, k_y)$ on the i-th layer and the two parameters Δ_S and Δ_D define the tunneling strength between the top and bottom surface of the same topological insulator layer and of neighbouring layers. One can show that the spectrum of the Hamiltonian is given by

$$\epsilon_\pm(\mathbf{k}) = \pm \sqrt{v_F^2 k_\perp^2 + \Delta^2(k_z)}, \tag{4}$$

where $\Delta(k_z) = \sqrt{\Delta_S^2 + \Delta_D^2 + 2\Delta_S \Delta_D \cos(k_z d)}$ and d is the superlattice period along z (given by the sum of the thicknesses of the two layers). This formula shows that the band structure has a gap, except if $\Delta_S = +\Delta_D$ or $\Delta_S = -\Delta_D$, in which cases the system is a semimetal, with Dirac points located at $k_z = \pi/d$ and $k_z = 0$, respectively. However, the latter phase corresponds to a critical point between a topological and a normal insulator and is, strictly speaking, unstable. Any deviation from $\Delta_S = \Delta_D$ or $\Delta_S = -\Delta_D$ (due, for instance, to strain or pressure effects) eliminates the degenerate Dirac point and induces a gap. The phase can be made stable by breaking inversion symmetry or time-reversal symmetry, which leads to separated Dirac points in momentum space [10].

Tominaga *et al.* [6] assumed that model 1 holds true and, thus, the inverted Petrov sequence corresponds to the high-resistance RESET state, whereas the ferroelectric configuration is the low-resistance SET state. They claimed that a moderate electric field should open a gap in the inverted Petrov state, resulting in a decrease in conductivity. A sufficiently strong field should instead cause the transition to the ferroelectric state and, thus, an abrupt increase in conductivity. Such non-ohmic behaviour is in fact observed in IPCM devices, as shown in Fig. 5.

Recently, it has been shown that stable topological 3-dimensional Dirac semimetals exist even in the presence of both inversion symmetry and time-reversal symmetry, if additional uniaxial rotational symmetries are present [30]. In this work, a complete classification of such topologi-

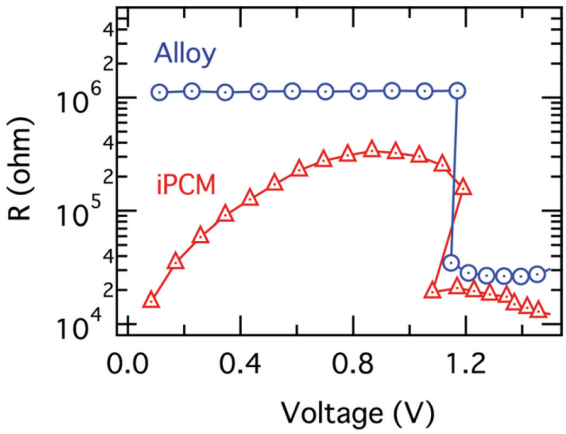


Fig. 5: *Experimental R-V characteristics obtained for a $[(GeTe)_2\ (Sb_2Te_3)_4]_8$ IPCM superlattice (red curve). Strong deviations from the Ohmic behaviour are observed during the switching of the superlattice. According to Ref. [6], the increase in resistivity for small voltages originates from the opening of a band gap in the inverted Petrov RESET state. The abrupt decrease in resistivity at large voltages is due to the transition to the ferroelectric SET state. The blue curve corresponds to a device using composite GeSbTe, which instead exhibits perfectly Ohmic behaviour. Reprinted from Ref. [6].*

cal phases has been carried out. Two distinct classes of topological semimetals have been shown to exist. The first class displays a single Dirac point at a time-reversal invariant momentum on the rotation axis. The second class exhibits a pair of Dirac points (created by band inversion), which lie on the rotation axis. It would be extremely interesting to investigate the relevance of these newly discovered topological states to $(GeTe)_2$-Sb_2Te_3 IPCM. Nevertheless, it is crucial to know the exact structural properties of the system before embarking on a study of the topology of the electronic structure. As we discuss in the next section, very recent experiments seem to suggest that none of the 4 phases discussed so far exactly describes the structure of these superlattices.

4 Latest developments

Very recent scanning transmission electron microscopy experiments on GeTe-Sb_2Te_3 superlattices [31] seem to challenge the current understanding of the structural properties of ICPM. In this work, cross-sectional high-angle annular dark-field (HAADF) imaging has been employed to characterize $[GeTe(1nm)\text{-}Sb_2Te_3(3nm)]_{15}$ superlattices artificially grown on passivated Si(111) at 230 °C using molecular beam epitaxy. The most stable structure has been found to consist of Sb_2Te_3 and rhombohedral GeSbTe (not GeTe!) building blocks (see Fig. 6). The thin layers of Sb_2Te_3 and GeSbTe are bonded via van der Waals interactions and thus form a van der Waals heterostructure. As a result of this reconfiguration of the superlattice, many defects are present, including stacking faults and layering disorder. Momand *et al.* have also tried to introduce sharp interfaces between GeTe and Sb_2Te_3 by using growth interrupts but, even in this case, the formation of rhombohedral GeSbTe has been observed [31].

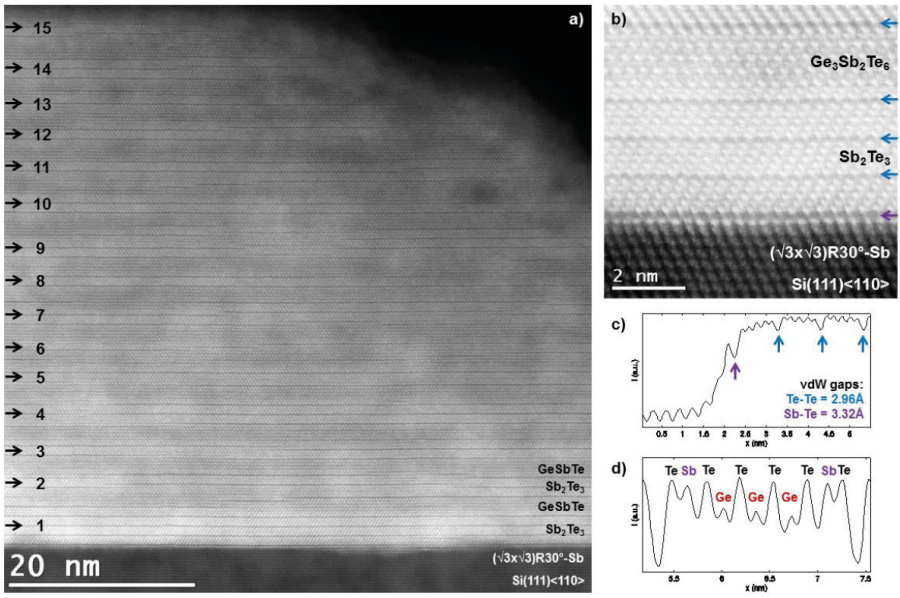


Fig. 6: *HAADF scanning transmission electron measurements on the as-deposited [GeTe(1nm)-Sb_2Te_3(3nm)]$_{15}$ superlattice grown by molecular beam epitaxy [31]. (a) Overview micrograph of the superlattice. (b) Close-up of the Si(111)-Sb-Sb_2Te_3 interface and the GeSbTe layered structure. (c) Intensity linescan of the Si(111)-Sb-Sb_2Te_3 interface in Fig. b. (d) Intensity linescan of the GeSbTe layer in Fig. b, which shows that its stoichiometry is $Ge_3Sb_2Te_6$. Reprinted from Ref. [31] - Published by The Royal Society of Chemistry.*

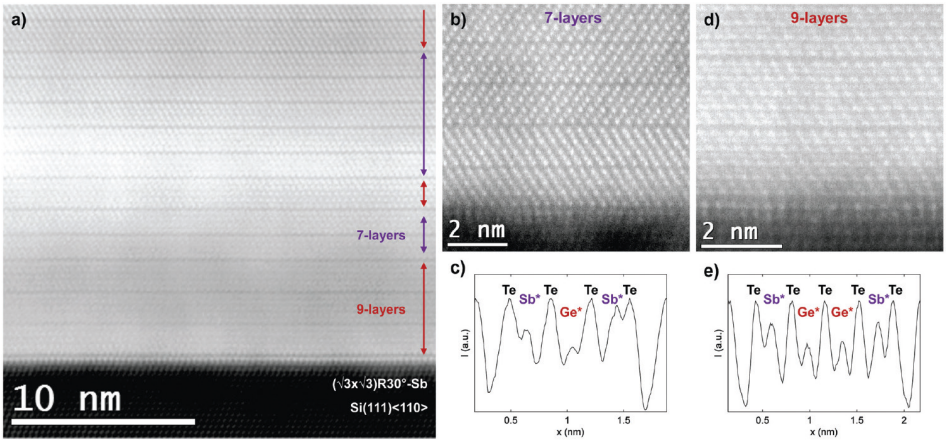


Fig. 7: *HAADF scanning transmission electron measurements on the [GeTe(1nm)-Sb_2Te_3(3nm)]$_{15}$ superlattice grown by molecular beam epitaxy, after annealing at* 400 °*C for 30 minutes [31]. (a) Overview micrograph showing that the superlattice has transformed to rhombohedral GeSbTe upon annealing. The latter consists of 7- and 9-layered van der Waals blocks. (b-c) Close-up of a region consisting of 7-layered van der Waals blocks and intensity linescan of one such layer. (d-e) Close-up of a region consisting of 9-layered blocks and intensity linescan of one such layer. The asterisks in Fig. c and e indicate intermixing between Ge and Sb atomic planes. Reprinted from Ref. [31] - Published by The Royal Society of Chemistry.*

These findings have been qualitatively explained in terms of the different bonding dimensionality of bulk GeTe and Sb_2Te_3 [31]. Sb_2Te_3 and GeTe are two-dimensionally and three-dimensionally bonded solids. Moreover, quintuple layers of Sb_2Te_3 are chemically passive (which results in van der Waals interactions between neighbouring Te layers in bulk Sb_2Te_3) and the formation of bonds with GeTe is energetically unfavorable. According to the authors, this property explains why, at zero temperature, the Kooi structure (in which GeTe is intercalated within the Sb_2Te_3 block, see subsection 3.1) is more favorable than the other sequences. Furthermore, the Kooi nonuple layer (Te-Sb-Te-Ge-Te-Ge-Te-Sb-Te) is chemically passive as well and can thus bind with Sb_2Te_3 layers by weak van der Waals interactions. This implies that the van der Waals gap is always formed after the -Te-Sb-Te termination of the stack. In the metastable superlattices grown by the authors, 1-2 quintuple layers of Sb_2Te_3 alternate with GeSbTe Kooi-like layers with intercalated GeTe. More precisely, different GeSbTe layered systems can form in between Sb_2Te_3 layers. 7-, 9-, 11- and 13-layered GeSbTe systems have been observed, corresponding to $GeSb_2Te_4$, $Ge_2Sb_2Te_5$, $Ge_3Sb_2Te_6$ and $Ge_4Sb_2Te_7$, respectively [31].

Momand *et al.* have also shown that, by annealing the superlattices at 400 °C for 30 minutes, a) there is a transition to the bulk rhombohedral GeSbTe structure, which is the thermodynamically stable phase (see Fig. 7) and b) a strong tendency for the Ge and Sb layers to intermix is observed, indicating that configurational entropy effects play an important role (as already mentioned in subsection 3.1).

It is very important to stress that the superlattice structures of Ref. [31] exhibit the typical phase-change behaviour of IPCM: in particular, the switching power of memory cells containing

these superlattices is an order of magnitude lower than that of GeSbTe cells. The fact that the structural properties of these superlattices are in disagreement with the previously proposed models casts doubts on the switching mechanisms discussed so far. It also raises the question whether the switching process involves a transition between two crystalline states or it is in fact due to amorphous-crystalline transitions of the thin GeSbTe sublayers. In the latter case, the reduced switching energy may be due to interfacial and/or strain energy effects [31]. In this respect, it was demonstrated that the energy of crystalline-amorphous interfaces can be lower than the energy of the crystalline-crystalline counterparts under certain conditions [32]. Moreover, strain could lower the amorphization energy for the rhombohedral GeSbTe layers, and template growth of GeSbTe from the (crystalline) Sb_2Te_3 matrix could result in higher growth speeds.

In conclusions, these findings indicate that the switching models which have been proposed so far may not describe the actual phase-change mechanisms occurring in IPCM. An atomistic understanding of the relevant processes is obviously crucial to further improve the performance and functionality of IPCM-based devices. Hence, there is pressing need for further experimental and theoretical investigations to elucidate the properties of these extraordinary materials.

References

[1] M. Wuttig and N. Yamada, Nature Mater. **6**, 824 (2007).

[2] T. C. Chong, L. P. Shi, X. Q. Wei, R. Zhao, H. K. Lee, P. Yang and A.Y. Du, Phys. Rev. Lett. **100**, 136101 (2008).

[3] J. Tominaga, P. Fons, A. Kolobov, T. Shima, T. C. Chong, R. Zhao, H. K. Lee and L. Shi, Jpn. J. Appl. Phys. **47**, 5763 (2008).

[4] R. E. Simpson, P. Fons, A. V. Kolobov, T. Fukaya, M. Krbal, T. Yagi and J. Tominaga, Nature Nanotechnol. **6**, 501 (2011).

[5] T. Matsunaga and N. Yamada, Phys. Rev. B **69**, 104111 (2004).

[6] J. Tominaga, A. V. Kolobov, P. Fons, T. Nakano and S. Murakami, Adv. Mater. Interfaces **1**, 1300027 (2014).

[7] T. Egami, K. Johguchi, S. Yamazaki and K. Takeuchi, Jpn. J. Appl. Phys. **53**, 04ED02 (2014).

[8] D. Bang, H. Awano, J. Tominaga, A. V. Kolobov, P. Fons, Y. Saito, K. Makino, T. Nakano, M. Hase, Y. Takagaki, A. Giussani, R. Calarco and S. Murakami, Sci. Rep. **4**, 5727 (2014).

[9] M. Z. Hasan and C. L. Kane, Rev. Mod. Phys. **82**, 3045 (2010).

[10] A. A. Burkov and L. Balents, Phys. Rev. Lett. **107**, 127205 (2011).

[11] K. Makino, J. Tominaga, A. V. Kolobov, P. Fons and M. Hase, Appl. Phys. Lett. **101**, 232101 (2012).

[12] T. Ohyanagi, N. Takaura, M. Kitamura, M. Tai, M. Kinoshita, K. Akita, T. Morikawa and J. Tominaga, Jpn. J. Appl. Phys. **52**, 05FF01 (2013).

[13] K. Makino, Y. Saito, P. Fons, A. V. Kolobov, T. Nakano, J. Tominaga and M. Hase, Appl. Phys. Lett. **105**, 151902 (2014).

[14] T. Ohyanagi, M. Kitamura, M. Araidai, S. Kato, N. Takaura and K. Shiraishi, Appl. Phys. Lett. **104**, 252106 (2014).

[15] X. Yu and J. Robertson, Sci. Rep. **5**, 12612 (2015).

[16] N. Govind, M. Petersen, G. Fitzgerald, D. King-Smith, and J. Andzelm, Comput. Mater. Sci. **28**, 250 (2003).

[17] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[18] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[19] B. J. Kooi and T. M. J. De Hosson, J. Appl. Phys. **92**, 3584 (2002).

[20] I. I. Petrov, R. M. Imanov and Z. G. Pinsker, Sov. Phys.-Crystallogr. **13**, 339 (1968).

[21] K. Shportko, S. Kremers, M. Woda, D. Lencer, J. Robertson, and M. Wuttig, Nat. Mater. **7**, 653 (2008).

[22] D. Lencer, M. Salinga, B. Grabowski, T. Hickel, J. Neugebauer, and M. Wuttig, Nat. Mater. **7**, 972 (2008).

[23] Z. Sun, J. Zhou and R. Ahuja, Phys. Rev. Lett. **96**, 055507 (2006).

[24] G. C. Sosso, S. Caravati, C. Gatti, S. Assoni, and M. Bernasconi, J. Phys.: Condens. Matter **21**, 245401 (2009).

[25] W. Zhang, A. Thiess, P. Zalden, R. Zeller, P. H. Dederichs, J-Y. Raty, M.Wuttig, S. Blügel and R. Mazzarello, Nature Mater. **11**, 952 (2012).

[26] L. Fu, C. L. Kane and E. J. Mele, Phys. Rev. Lett. **98**, 106803 (2007).

[27] D. Hsieh, Y. Xia, D. Qian, L. Wray, F. Meier, J. H. Dil, J. Osterwalder, L. Patthey, A.V. Fedorov, H. Lin, A. Bansil, D. Grauer, Y. S. Hor, R. J. Cava, and M. Z. Hasan, Phys. Rev. Lett. **103**, 146401 (2009).

[28] J. Kim, J. Kim, and S.-H. Jhi, Phys. Rev. B **82**, 201312 (2010).

[29] J. Kim, J. Kim, K.-S. Kim, and S.-H. Jhi, Phys. Rev. Lett. **109**, 146601 (2012).

[30] B.-J. Yang and N. Nagaosa, Nat. Comm. **5**, 4898 (2014).

[31] J. Momand, R. Wang, J. E. Boschker, M. A. Verheijen, R. Calarco, and B. J. Kooi, Nanoscale **7**, 19136 (2015).

[32] R. Benedictus, A. Böttger, and E. J. Mittemeijer, Phys. Rev. B **54**, 9109 (1996).

D9 Memristive Tunneling Devices: From Device Principles to Neuromorphic Applications

M. Ziegler, A. Petraru, R. Soni, and H. Kohlstedt
Nanoelektronik, Technische Fakultät der
Christian-Albrechts-Universität zu Kiel, Germany

Contents

1 Introduction **2**

2 Current Transport through Energy-Barriers **2**
2.1 Formation of Barriers 3
2.2 Electron Tunneling 4
2.3 Thermionic Emission Theory 6

3 Memristive Tunneling Devices **7**
3.1 Floating Gate Transistors as memristive Devices: MemFlash Cell 8
3.2 Interface-based Memristive Devices: A memristive Tunnel Junction 13
3.3 Ferroelectric Tunneling Junctions 16

4 Applications in Neuromorphic Systems **19**
4.1 Memristive *Hebbian* Plasticity 19
4.2 Device Requirements for the Emulation of *Hebbian* Plasticity 20

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

Even if it has been known for more than *50* years that metal oxides exhibit a more or less abrupt transition from an insulating to a conducting state under certain experimental conditions [1], named *resistive switching*, a revival has been experienced in the 1990s, when researchers start to think about new computer memory concepts. A remarkable second increase of scientific interest in resistive switching began in 2008, when Strukov *et al.* [2] linked resistive switching to the *memristor* postulated by Chua in 1971 [3]. This particularly allows describing resistive switching devices as a memristive system which is defined by the following set of equations:

$$\begin{aligned} i(t) &= g(x,u,t)\,u(t) \\ \frac{dx}{dt} &= f(x,u,t) \end{aligned} \tag{1.1}$$

Here, *x* describes the state of the system, while *i(t)* and *u(t)* can be identified by the current and the voltage of the device. The response function g corresponds to the device conductance, named for a memristive system also *memductance*. Further, f is a continuous function, which describes the dynamics of the resistive switching process.

Memristive devices are considered today as potential candidates for future non-volatile data storage technologies and as key devices in a variety of electronic circuits, as for example field programmable gate arrays (FPGAs) or artificial neural networks (ANN). Due to their simple two terminal capacitor-like layer sequence (metal-insulator-metal), memristive devices might overcome technical and physical scaling limits of modern semiconductor devices [4].

While the technical realization of memristive devices can be rather simple, the underlying physical mechanisms are very diverse and complex [5]. This holds in particular for devices involving ionic conductance mechanisms. The majority of memristive devices involve random creation of one or more conductive filaments, resulting in a poor switching reproducibility and a high device-to-device variability [6]. In this regard, in this chapter we like to focus on memristive tunneling devices, which may overcome these restrictions. A common feature of those devices is that the resistance changes due to modification of energy barriers for the electron transport, and the formation of conductive filaments is particularly avoided.

This chapter is structured in three parts: At first we like to briefly show the main mechanisms contributing to current conductance through energy barriers. Thereafter in chapter 3 different memristive device concepts based on electron tunneling are presented. Finally, to show some application field of memristive tunneling devices, their use in *neuromorphic* systems are briefly overviewed in chapter 4. In this field memristive devices are used as technical substitute of chemical synapses in artificial neural networks.

2 Current Transport through Energy-Barriers

Interface processes are crucial for the conductance properties of electronic devices: If a metal is connected with an insulator or a semiconductor an *energy barrier* is formed, which is responsible for the device performance. Particularly, for memristive tunneling devices the energy barriers play a key role. In this section we like to briefly overview over the energy barrier concepts by mainly focusing on electronic transport. The here discussed theoretical models are adapted from Ref. [7], where the reader is referred to for further details.

2.1 Formation of Barriers

In Fig. 2.1 simplified *energy-band* diagrams for a metal (a) and a semiconductor (b) are depicted. The quantity $q\phi_m$ denotes the *work function* of a metal, which is defined as the energy difference between the *Fermi* level E_F and the *vacuum* level. In a semiconductor, see Fig. 2.1(b), the work function is equal to $q(\chi+\phi_n)$, where $q\chi$ is the electron affinity, which defines the energy required to move an electron from the bottom of the *conductance band* E_C to the vacuum level, and $q\phi_n$ is the energy difference between E_C and E_F.[7] For the devices discussed in chapter 3, two structures are of particular interest: two metals separated by an ultra-thin insulator and a metal in contact with a semiconductor, which form a *tunneling* and *Schottky* barrier, respectively. In the following both energy barrier types are briefly overviewed.

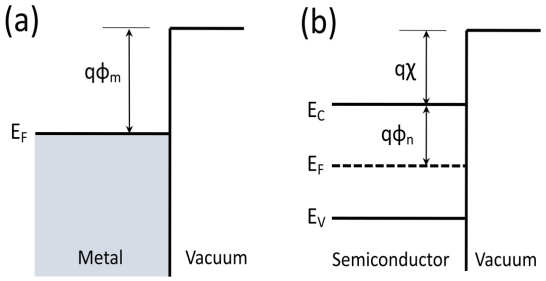


Fig. 2.1: *Energy-band diagrams of a metal (a) and a semiconductor (b). While $q\phi_m$ denotes the work function of a metal, $q(\chi+\phi_n)$ is equal to the work function in a semiconductor. Adapted from [7].*

Tunneling Barrier

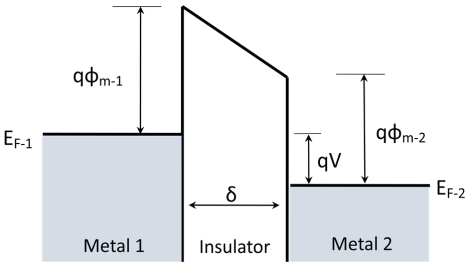


Fig. 2.2: *Schematic energy-band diagram of a tunneling contact under bias voltage application V: Two metals with the respective work functions ϕ_{m-1} and ϕ_{m-2} are separated by a thin insulator with a thickness δ.*

In Fig. 2.2 an energy-band diagram of a metal-insulator-metal contact is presented. If the width of the insulator δ is in the range of an electron wave length, electron can tunnel from metal 1 two metal 2 through the energy barrier. In a simplistic model the effective tunneling barrier can be described in the framework of *Simmons tunneling model* [8], which underlying the assumptions, that the potential is spatially averaged and varies linearly with space and applied voltage. In particular, the model is based on averaging the potential-thickness profile of the tunnel barrier, resulting in a single characteristic parameter, the averaged energy barrier:

$$q\Phi(\mathrm{V}) = \frac{q(\phi_{m-1} + \phi_{m-2})}{2} + \frac{qV}{2} \tag{2.1}$$

It is worth to mention that Eq. 2.1 represents a special case of the general Simmons model, wherefore two very stringent assumptions have been made: (i) the potential profile is spatially linear, meaning that it has a trapezoidal shape, and (ii) the applied bias voltage V adds linearly to the potential profile.

Schottky Barrier

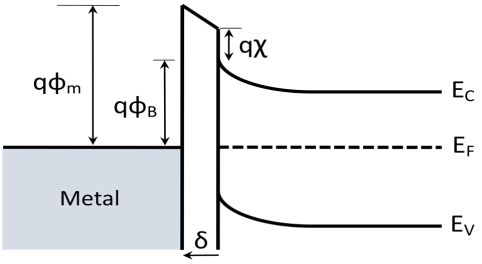


Fig. 2.3: *Schematic energy-band diagram of metal n-type semiconductor contact. Here $q\phi_m$ denotes the work function of the metal, while $q\phi_B$ and $q\chi$ are the Schottky barrier height and the electron affinity, respectively. Adapted from [7].*

In Fig. 2.3 a simplified energy-band diagram of a metal *n-type* semiconductor contact is shown, which has been adapted from [7]. As first supposed by *Schottky* in 1938, an energy barrier in metal-semiconductor contacts arises from stable space charges in the semiconductor without the presence of a chemical layer. Hence, an energy barrier $q\phi_B$ (also named as *Schottky barrier*) is formed if a metal is in contact with a semiconductor. As sketched in Fig. 2.3 if the gap δ between metal and semiconductor decreases, the electrical field in the gap increases and builds-up a negative charge at the metal surface. This negative charge has to be compensated by a positive charge in the semiconductor. In the ideal case δ would be zero, where the gap becomes transparent to electrons and the energy barrier height for an *n*-type semiconductor is given by

$$q\phi_B = q(\phi_m - \chi). \tag{2.2}$$

Hence in an ideal metal-semiconductor contact the height of the energy barrier is given by the difference between the metal work function and the electron affinity. However, in reality (as we will discuss it in chapter 3) the ideal conditions assumed for Eq. 2.2 are never satisfied. In reality the Schottky barrier height is mainly modified by interfacial layers ($\delta \neq 0$), interface states, and image force lowering.

2.2 Electron Tunneling

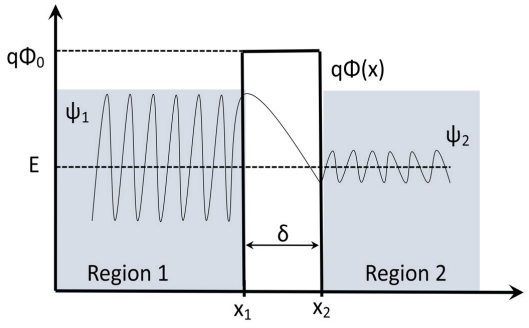


Fig. 2.4: *Principle of electron tunneling through a rectangular energy barrier with thickness δ. ψ_1 and ψ_2 are the electron wave functions in region 1 and 2, respectively. Adapted from [7].*

The concept of *elastic electron tunneling* through an energy barrier is sketched in Fig. 2.4. In particular, tunneling is a quantum mechanical phenomenon in which an electron can be described by its wave function ψ. This, in fact allows an electron guarding an energy E well below the height of the energy barrier Φ to tunnel from region 1 into region 2. Therefore, the barrier width δ must be sufficiently small, so that the electron wave function ψ_1 from region 1 can be recaptured in region 2. As a result, the electron ψ_2 has a decreased amplitude compared to ψ_1 but keeps its energy E (cf. Fig. 2.4), which means that the tunneling process is *elastic*. By using the wave functions ψ_1 and ψ_2 the probability of an electron with an energy E to be transmitted through the barrier can be calculated by

$$T = \frac{|\psi_2|^2}{|\psi_1|^2}. \tag{2.3}$$

This ratio is called *transmission probability*. In order to calculate T, the electron wave function can be expressed as a planar wave of the form $exp(\pm ikx)$, where k is the wave vector and x the spatial position of ψ. By using $E = (\hbar k)^2/(2m^*) + q\Phi(x)$ (m^* is the effective electron mass and $\hbar$ *Plancks* constant) and if we assume that the potential Φ (x) does not vary rapidly, according to WKB (Wentzel-Kramers-Brillouin) approximation, Eq. 2.3 can be rewritten as

$$T = \exp\left\{-2\int_{x1}^{x2}\sqrt{\frac{2m^*}{\hbar^2}[q\Phi(x) - E]}\,dx\right\}. \tag{2.4a}$$

If for the potential barrier Φ (x) in Eq. 2. 4 the averaged barrier height Eq. 2.1 is used, the transmission probability can be simplified to

$$T(V, E, \delta) = \exp\left\{-2\frac{\sqrt{2m^*}}{\hbar^2}\delta\sqrt{\frac{q(\phi_{m-1} + \phi_{m-2})}{2} + \frac{qV}{2} - E}\right\}. \tag{2.4b}$$

Therefore, the probability of electron tunneling is restricted to the physical parameters of the energy barrier, which allows modifying the current transport in respect to a variation of these parameters. This important characteristic can be used to realize memristive devices, as it will be presented in chapter 3.

Together with Eq. 2.4a, the tunneling current I, can be calculated from the product of the number of available electrons in the originating region 1 and the number of empty states in the destination region 2:

$$I = \frac{qm^*}{2\pi^2\hbar^3}\int_{-\infty}^{+\infty} N_1 f_1 \cdot N_2(1 - f_2) \cdot T\,dE \tag{2.5}$$

where f_1, f_2 and N_1, and N_2 are the Fermi-Dirac distributions and densities of states in the corresponding regions, respectively.

A particular important assumption made for the derivation of the presented tunneling model is that the applied voltage V is much smaller than the height of the energy barrier. In particular, for higher electrical voltages this tunnel model is not more applicable. However, in particular, in the field semiconductor technologies tunnel processes at elevated voltages are important, since those mechanisms are used for *non- volatile* memory devices to store and erase information. The principle of electron tunneling at higher voltages can be obtained from the *Fowler-Nordheim* tunneling model, which assumes that due to the high applied voltage V the barrier is deformed into a triangularly shape and the electron transmission probability increas-

es with increasing the applied electrical field $\mathcal{E} = V/\delta$ through the barrier. Within this model the tunneling current reads

$$I_{FN} = A_j A_{FN} \mathcal{E}^2 \exp\left(-\frac{B_{FN}}{\mathcal{E}}\right). \tag{2.6}$$

Here, is A_j the tunneling area and A_{FN}, B_{FN} are constants which are given by $A_{FN} = q^3 m_0/(8\pi hm^*\Phi)$ and $B_{FN} = 8\pi\,(2m^*\Phi^3)^{1/2}/(3qh)$.

2.3 Thermionic Emission Theory

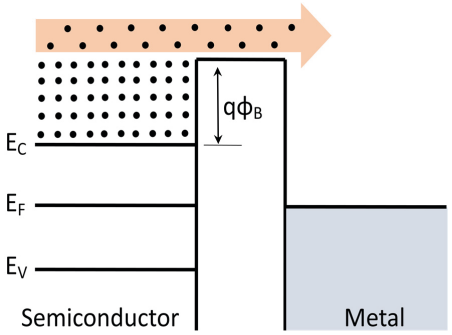


Fig. 2.5: *Schematic drawing of the thermionic emission of electrons over the energy barrier of a Schottky contact. Adapted from [7].*

The current conduction mechanism in metal-semiconductor contacts can be described by the *thermionic emission theory*. This theory is based on three major assumptions: (1) the barrier height $q\Phi_B$ is assumed to be much larger than kT; (2) drift diffusion effects within the barrier layer are neglected; and (3) the energy barrier is not affect by the image force. Thus, the current flow depends solely on the barrier height, as depicted in Fig. 2.5 (note that the explicit form of the barrier plays no role). Therefore, the current density of the electrons flowing from the semiconductor to the metal (named *forward current*) $J_{S\text{-}M}$ can be derived from the number of electrons above the energy barrier $q\Phi_B$, which reads

$$n = N_C \exp\left(-\frac{q(\Phi_B - V)}{kT}\right), \tag{2.7}$$

with N_C the effective density of states in the conductance band of the semiconductor. By using the current relation for a random motion of carriers within a *Maxwellian* distribution of velocities v_a, which is given by $J = nq/4$ with $v_a^2 = 8kT/(\pi m^*)$, we obtain with Eq 2.7

$$J_{s-M} = A^* T^2 \exp\left(-\frac{q(\Phi_B - V)}{kT}\right), \tag{2.8}$$

for the forward current density. Here A^* is the Richardson constant, which is given by $A^* = 4\pi q m^* k^2/h^3$. Moreover, the electron transport from the metal into the semiconductor can assumed to be unaffected by the applied voltage, i.e. V=0, so that the total current density, i.e. the sum of $J_{S\text{-}M}$ and $J_{M\text{-}S}$, reads

$$J_{s-M} = A^* T^2 \exp\left(-\frac{q(\Phi_B)}{kT}\right)\left[\exp\left(\frac{qV}{kT}\right) - 1\right]. \tag{2.9}$$

In reality, the current transport can be modified by quantum-mechanical effects within the barrier region. Crowell and Sze have extend the thermionic emission model by quantum mechanical tunneling and reflection and found that these effects end up in a reduced effective Richardson constant A^{**} [7].

3 Memristive Tunneling Devices

The main idea of memristive tunneling devices is to change the electronic conduction of the device by using the electron tunneling mechanism. While the electronic tunneling process plays a central role in those devices, the underlying physical principles and device concepts can be rather different. In Fig. 3.1 several so far proposed memristive tunneling concepts which based-on mixed ionic-electronic, magnetic, and purely electronic resistance switching mechanisms are shown. Regarding this diversity of switching mechanisms it is worth to group memristive tunneling devices into two classes: while the first class of devices uses electron tunneling to change a reference potential (i.e. a gate potential) which controls the device conductance. The conductance of the second class of tunneling devices is controlled by variations of energies barriers induced by the applied electrical field.

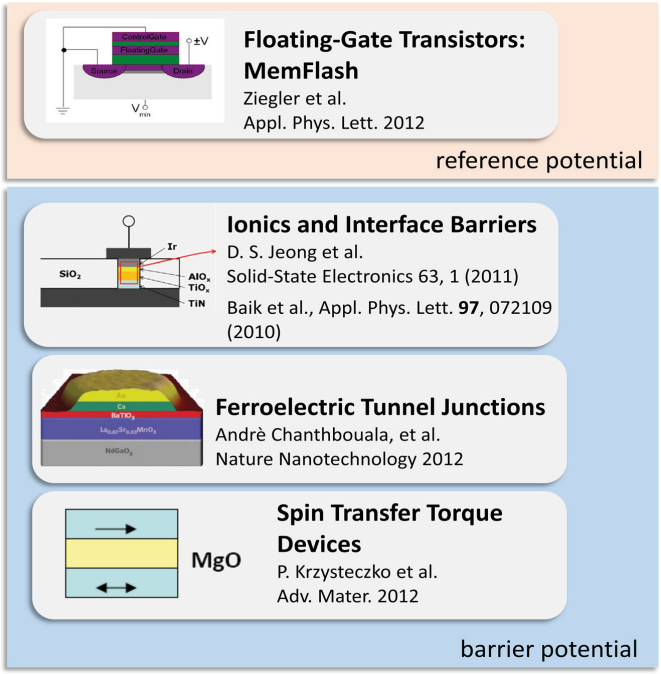


Fig. 3.1: *Overview of different memristive tunneling device, which are grouped in two classes. While the first class of devices uses electron tunneling to change a reference potential (i.e. a gate potential) in order to control the device conductance, the conductance of the second class of tunneling devices is controlled by variations of energies barriers induced by the applied electrical field.*

In this section we like to discuss three distinct memristive tunneling device concepts: The first memristive device concept that we would like to present here belongs to the first class of devices referred in Fig. 3.1 and it is based-on state-of-the art floating-gate transistors, named *MemFlash*-cells (Sec. 3.1). Thereafter, two memristive tunneling devices which belong to the second class of tunneling devices are presented. While in Sec. 3.2 the possibility to use mobile ions in order to change energy barrier properties in memristive devices isdiscussed, in Sec. 3.3 *ferroelectric* tunnel junctions are presented for which ferroelectric materials are employed as tunnel barriers.

3.1 Floating Gate Transistors as memristive Devices: MemFlash Cell

Floating-gate transistors are count to the major memory devices for non-volatile storage technologies, such as Flash-cells. In particular, a floating gate transistor is obtained when the gate electrode of a conventional MOS-FET (Metal-Oxide-Semiconductor-Field Emission Transistor) is modified to incorporate an additional metal-insulator sandwich, i.e. floating-gate. The floating gate allows being semi permanent charged so that a memory effect is implemented [7].

However, floating-gate transistors are three-terminal devices with separated read and write cycles and therefore they cannot be regarded as memristive system at a first glance according to Eq. 1.1: For a memristive device simultaneous read/write functionality is required. In the following section we will show that a particular wiring scheme, however, allows a memristive operation mode of the floating gate transistors, named *MemFlash-cell*. We show evidence that the *MemFlash* can be considered as a potential substitute for any memristive device (especially for reconfigurable logic, cross-bar arrays, and neuromorphic circuits) and it is basically compatible with current Si-fabrication technology. The *MemFlash* concept presented here is adapted from [9, 10], where the reader is referred to for further details.

Device Principle

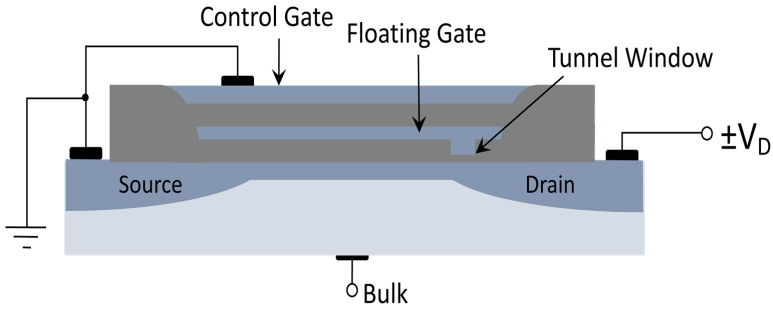


Fig. 3.2: *Schematic drawing of a MemFlash-cell. The diode wiring scheme enables a memristive operation mode of the device, such as that the resistance changes depending on the charge flow through the device. Regarding this, the CG and the source (S) terminals of the device are set to the common ground, while a bipolar voltage is supplied through the drain terminal (D). In order to avoid a short cut between the drain and the substrate bulk, the bulk terminal has been set to the minimal voltage of the bipolar voltage supply through the drain terminal.*

The wiring scheme that enables the memristive operation mode of a floating-gate transistor is depicted in Fig. 3.2. As a floating gate transistor a conventional EEPROM cell (electrical erasable programmable read only memory) cell has been employed in [9], which is a typical storage cell of conventional Flash memories. In order to ensure the memristive operation mode, the external accessible terminals source (S) and control gate (CG) are connected to the common ground potential of the circuitry, while a bipolar voltage supply is connected to the drain terminal (D). Further, the bulk terminal (B) is set to the minimum voltage of the bipolar voltage supply in order to guarantee the formation of a conductive channel between source and drain and to avoid a short-circuit fault to the source. The therewith obtained current-voltage characteristics of the two terminal circuitry, using a single EEPROM cell are depicted

in Fig. 3.3. It shows the typical *pinched hysteretic* loop of a memristive device. Therefore, the bipolar voltage was linear ramped between ±*12* V, while the drain current was recorded simultaneously.

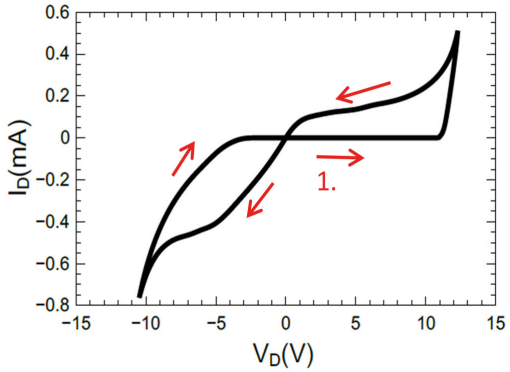


Fig. 3.3: *Typical current-voltage characteristic of a MemFlash cell. The red arrows indicating the voltage sweep direction. Adapted from [9]*

The working principle of the MemFlash cell can be understood in the framework of the *Fowler-Nordheim* tunneling process: At positive drain voltages electrons are tunneling from the floating gate through the tunneling oxide (located at the drain electrode and highlighted in Fig. 3.2) into the *MOS-FET* channel and vice versa for negative drain voltages. Therefore the floating gate charge is modified during the voltage sweep, which itself changes the overall device resistance.

Capacitive device model

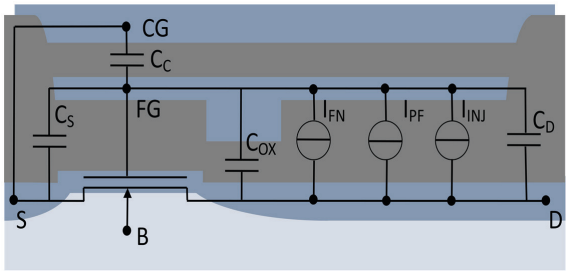


Fig. 3.4: *Capacitive device model of a MemFlash-cell. Adapted from [10]*

To get further insight into the device mechanism, a simple capacitive model can be employed for the device description, as depicted in Fig. 3.4. Within this purely capacitive model, the floating gate potential is expressed as

$$V_{FG} = \frac{Q_{FG}}{C_T} + k_C V_C + k_D V_D + k_S V_S + k_B V_B \tag{3.1}$$

where Q_{FG} is the charge stored on the floating gate, V_C, V_D, V_S, and V_B are the potentials of the control gate, drain, source, and bulk terminal, respectively. Furthermore, k_C, k_D, k_S, and k_B are the respective coupling constants to the floating gate electrode, which are defined by $k_i = C_i/C_T$ ($i = C, D, S, B$) with the capacities C_C, C_D, C_S, C_B, and the total capacity $C_T = C_C + C_D + C_S + C_B$. In the memristive operation mode, a three terminal floating gate transistor is reconfigured to a two-terminal cell. Therefore, V_C and V_S are connected to the common ground

potential, i.e. $V_C = V_S = 0$. Further, if the floating gate coupling to the bulk is sufficiently small (i.e. $k_B \approx 0$), then Eq. 3.1 can be rewritten by

$$V_{FG} = \frac{Q_{FG}}{C_T} + k_D V_D . \tag{3.2}$$

Eq. 3.2 shows that the floating gate potential is determined by the drain voltage V_D and the history of V_D through the floating gate charge Q_{FG}, which can be calculated by

$$Q_{FG}(t_1) = Q_{FG}(t_0) + \int_{t0}^{t1} I_{FN}(V_{FG}(t), V_D(t))\, dt . \tag{3.3}$$

Here, I_{FN} is the *Fowler-Nordheim* tunneling current, as defined in Eq. 2.6.

While the *Fowler-Nordheim* tunneling current is mainly contributing to charging and discharging of the floating gate potential, additional current contribution may arise from localized defect states inside the tunneling oxide and from electrons with energies above the barrier height of the tunneling oxide barrier, so called hot electrons. According to Ref. [10] defect states inside the tunneling oxide can be approximated by the *Poole-Frenkel* current, which is given by

$$I_{PF} = \pm A_{tox} A_{PF} E_{tox} \exp\left(B_{PF} \sqrt{E_{tox}}\right). \tag{3.4}$$

Here, A_{PF} and B_{PF} are positive constants. The additional contribution to Q_{FG} arising from the injection of hot electrons can be approximated by

$$I_{inj} = \pm A_{tox} \mathrm{A}_{\mathrm{inj}} \exp\left(- \frac{\mathrm{B}_{\mathrm{inj}}}{\left(\mathrm{C}_{\mathrm{inj}} + \mathrm{V}_{\mathrm{FG}}\right)^2} + \mathrm{D}_{\mathrm{inj}}\, \mathrm{V}_{\mathrm{D}} \right). \tag{3.5}$$

Where A_{inj}, B_{inj}, C_{inj} and D_{inj} are positive constants, which have to be estimated by fitting the experimental data.

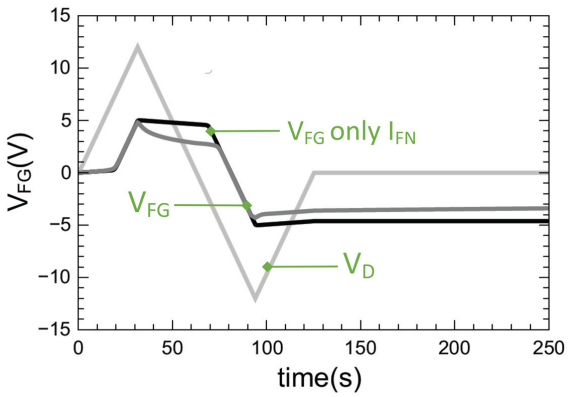


Fig. 3.5: *Voltage characteristic of the floating gate potential V_{FG} for a linear ramped drain voltage V_D calculated from Eq. 3.3. While the black curve shows the results if only Fowler-Nordheim tunneling is used, the gray line was obtained by taking into account additionally Pool-Franklin emission and hot electron injection.*

In Fig. 3.5 the voltage characteristic of the floating gate for a linear ramped drain voltage is shown which has been obtained from Eq. 3.3 and the parameters of Ref. [10]. The floating gate potential follows roughly the applied drain voltage, whereas the amplitude of V_{FG} is clearly reduced and temporally shifted compared to V_D. Further, from Fig. 3.5 it is also visible that the most important contribution for charging and discharging the floating gate originates

from the Fowler-Nordheim tunnelling (see black curve in compression to gray curve), while I_{PF} and I_{INJ} lead to a slide modification of the floating gate voltage characteristic, which becomes important if an accurate estimation of Q_{FG} is required.

In order to compare the discussed floating gate behaviour with experimental current-voltage characteristics of real MemFlash cells it is necessary to calculate the resulting drain current I_D of the underlying MOSFET, which can be described by the following set of equations [9]:

$$\begin{aligned} I_D &= \beta\,[(V_G - V_{th})\,V_D - \tfrac{1}{2}V_D^2]\,(1+\lambda\,V_D) && \text{for } V_G > V_{th} \text{ and } V_D < V_G - V_{th} \\ I_D &= \tfrac{1}{2}\,\beta\,(V_G - V_{th})^2\,(1+\lambda\,V_D) && \text{for } V_G > V_{th} \text{ and } V_D > V_G - V_{th} \\ I_D &= 0 && \text{for } V_G < V_{th} \end{aligned} \tag{3.7}$$

Here V_{th}, β, and λ are the threshold voltage, the transconductance, and the channel-length modulation parameter of the MOS transistor, respectively. V_G is the gate potential, which equals V_D-V_{FG} for $V_D < 0$ and $V_G = V_{FG}$ for $V_D > 0$.The therewith obtained current-voltage curve is shown in Fig. 3.6.

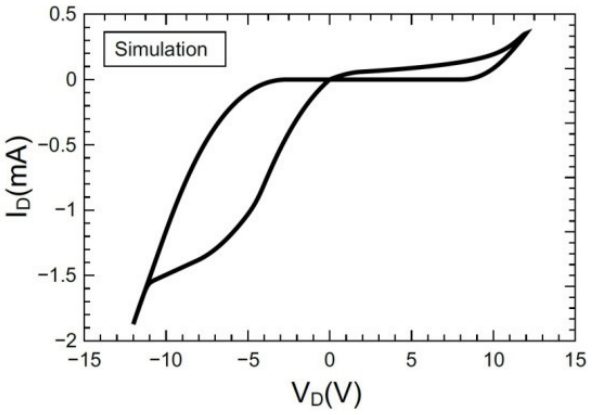


Fig. 3.6: *Calculated current-voltage characteristic of a MemFlash cell. Parameter of the MOSFET Model: $\lambda = 0.0625\ V^{-1}$, $\beta = 28.3\mu S/V$, and $V_{th} = 1.052\ V$.*

Scaling Perspective

The most striking disadvantage of a MemFlash cell compared to state-of-the art memristive devices is the power consumption. This can cause problems for usages in large circuits with many of these cells. One reason for this relative high power consumption is that the cells discussed here have a relatively large floating gate area of $A_{FG} = 84.98\ \mu m^2$. However, state-of-the-art EEPROM cells consist of floating gate areas in the nanometer range, which would reduce the source drain current by several orders of magnitudes. It is worth to mention, that the presented wiring scheme is not restricted to EEPROM cells and has been recently successfully applied to *quantum dot* floating gate transistors [11]. On the other side commercial EEPROM cells are designed for memory applications, where data retention of more than ten years is required. This, in fact, contains the disadvantages of relative thick tunneling oxides with the need of high programming voltages. At this respect, the gate oxide thickness represents a trade-off between low bias voltage (for low power consumption) and data storage retention times. In particular, for applications in neuromorphic, as it will be discussed in section 4, a retention time of ten years might not be necessary. Therefore, thinner floating gate oxide thickness could be accepted, leading to lower power consumption during programming.

In Fig. 3.7 the maximal drain voltages and minimal drain currents needed for charging and discharging the floating gate are plotted as function of the tunneling oxide thickness. For the presented data points the breakdown field strength of the tunneling oxide was assumed as maximal applicable electric field, which is 12 MV/cm for SiO_2. In particular, the minimal current was chosen to be displayed in Fig. 3.7, since it defines the maximal current flow through the device due to the asymmetry between the positive and negative voltage regime. From this rough approximation it is already visible that with a decreased tunneling oxide thickness the power consumption can be drastically decreased. Moreover, a reduced tunneling oxide thickness allows applying faster voltage sweeps, since the time constant of charging or discharging the floating gate is reduced.

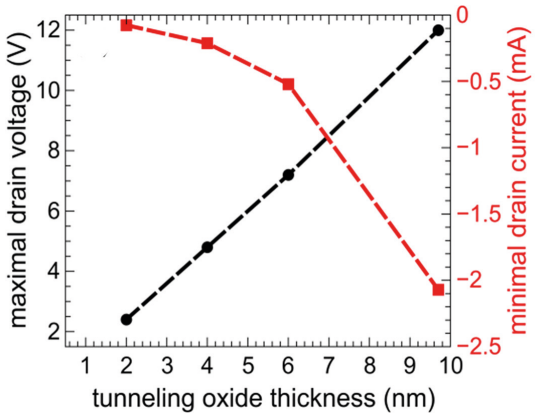


Fig. 3.7: *Estimation of power consumption of MemFlash cells with thinner tunneling oxides. For the estimation the breakdown field strength of SiO_2 has been used to calculate the maximal theoretical possible applicable drain voltage.*

However, the thickness of the tunneling oxide has to be a trade-off with the demand of data retention. For the MemFlash device, the resulting retention times for thinner tunneling oxides are shown in Fig. 3.8, which has been adapted from [10]. Therein V_{FG} is plotted as a function of time. It shows that retention times ranging from more than ten years up to several seconds are expected by varying the thickness of the tunneling oxide. Regarding this, a trade-off between the applications field, power consumption, and data retention is necessary;in neuromorphic circuits for example, tunneling oxide thicknesses in the range of 4 nm - 6 nm are of interest, which results in retention times ranging from one day to several months.

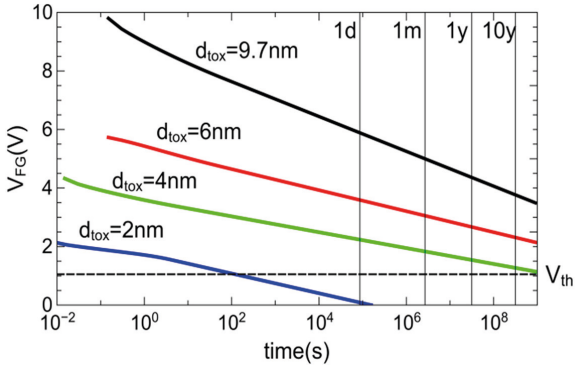


Fig. 3.8: *Simulation of the expected retention times for different gate tunneling oxide thickness. For the calculations, the floating gate potential V_{FG} is set to the maximal allowed potential for a given oxide thickness. Afterwards 45,000 integration steps are calculated and thereafter linear extrapolated. To guide the eyes vertical lines are added for different time values. Adapted from [10].*

3.2 Interface-based Memristive Devices: A memristive Tunnel Junction

Interface-based memristive devices can convince through their homogeneous switching characteristics [12-19]. While most of the investigated interfacial devices are oxide-metal junctions, where the resistive switching mechanism results from changes at a Schottky-like contact [14,20], a less common approach uses junctions consisting of a tunnel barrier and a memristive layer, where the change in resistance results from varying electron tunneling probability. [16,17,21,22] To explain the not completely understood resistance change in interface-based devices, the two usually considered models are depicted in Fig. 3.9. The first model is related to the concept of interfacial charges, which change the energy barrier height. In the second model, mobile ions within the memristive layer, lead to a change of the energy barrier of the interfacial layer. Moreover, beside this interface effects, contributions from the memristive layer itself (e.g. local chemical bounds, oxide phases, doping, local heating effects and so on) may additionally affect the resistive switching. This, in fact, makes the analysis of the underlying mechanism very complicated.

The idea used in Ref. [12] is to scale down the thickness of the memristive layer in the range of electron tunneling regime in order to avoid contributions of the memristive layer and to uncouple ionic from electronic effects. The additional use of a second barrier might restrict switching effects completely to interfacial contributions. In the following the concept of Ref. [12] is explained in detail.

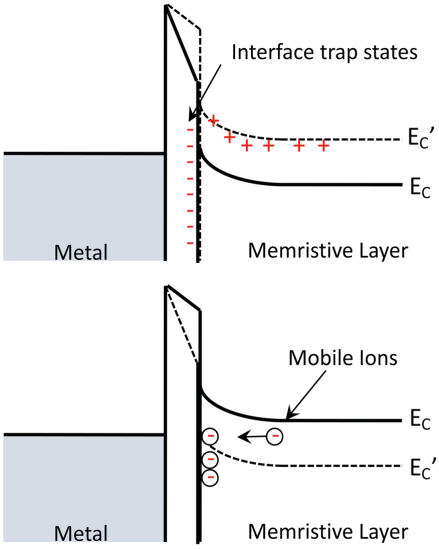


Fig. 3.9: *Cross-sectional view of interfacial resistive switching. Top: Trap states within the memristive leading to the filling and emptying of traps by injected electrons. Bottom: Negative ions can move inside the memristive layer.*

Device Structure and Resistive Switching Behavior

In [12] a double barrier memristive tunneling device was realized, in which the ultra-thin memristive layer is sandwiched between a tunnel barrier and a *Schottky*-like contact (see Fig. 3.10). The layer sequence of the device is Al/Al_2O_3/Nb_xO_y/Au, with a thickness of 1.3 nm for the Al_2O_3 tunnel barrier and 2.5 nm for the Nb_xO_y layer. A schematic energy band diagram is shown in Fig. 3.10, which shows a tunnel barrier at the Al_2O_3/Nb_xO_y- interface and a *Schott-*

ky-barrier at the Nb_xO_y/Au- interface. The device mechanisms based-on the mobile oxygen ions within the memristive Nb_xO_y-layer, which cause a decrease of the interfacial potential V_I by a down- shift of the interfacial energy band (conductance band) in Nb_xO_y (dashed line in Fig. 3.10). Regarding this down-shift, the effective tunnel distance d_{eff} and the energy barrier heights of the Al electrode ϕ_{Al}, and the Au contact ϕ_{Au} are decreased too. Hence, the overall device resistance is decreased.

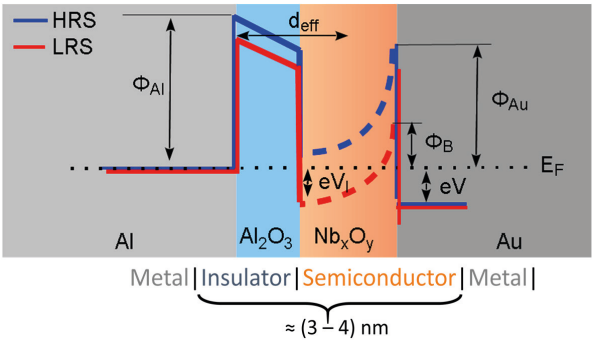


Fig. 3.10: *Cross-sectional view of a memristive tunneling device, which consist of metal-insulator-semiconductor-metal structure. Depending on the device resistance state, i.e. HRS (high resistance state) or LRS (low resistance state), the charge transport through the energy barriers is affected. Adapted from [12]*

A characteristic current-voltage curve of the memristive tunneling device of Ref. [12] is shown in Fig. 3.11. Therein, the absolute current density $|J|$ as function of the applied bias voltage is presented. Further, in the inset of Fig. 3.11 the area-resistance product vs. junction-area curve of the device is depicted, which has been measured at 0.5 *V*. In particular, the highly uniform current distribution for the LRS (low resistance state) and HRS (high resistance state) for areas ranging between 70 μm^2 and 2300 μm^2 indicates an interface based resistive switching mechanism.

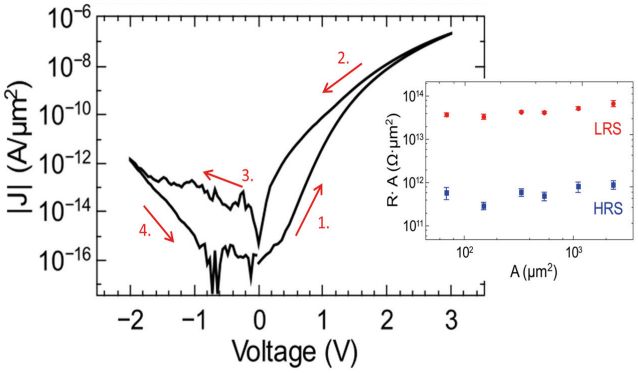


Fig. 3.11: *Absolute current density $|J|$ as function of the applied bias voltage of a memristive tunneling device. Inset: Area-resistance product vs. junction-area curve of the double barrier device measured at 0.5 V. Adapted from [12]*

Interface Energy Barriers Contributions

In order to get some more inside into the specific barrier contributions, an $Al/Al_2O_3/Nb_xO_y/Nb$ tunnel junction excluding the Schottky contact and an $Nb/Nb_xO_y/Au$ Schottky contact without the tunneling barrier has been compared in Ref. [12]. The therein obtained current-voltage characteristics are depicted in Fig. 3.12. While the memristive behavior is clearly visible for the Schottky-like contact, no change in the device resistance be-

havior is visible for a *"pure"* tunnel junction. In this reagard, the Schottky-like interface can be assumed as the active interface which is mainly responsible for resistance switching effect.

However, due to the ultra-thin memristive layer the two interfaces cannot be treated as separate entities and involve a very strong mutual interdependence. In order to discuss this aspect of the device in some more detail it is worth looking at the interfacial potential V_I (see the energy band-diagram in Fig. 3.10), i.e. the potential which influences both, the tunnel and the Schottky energy barrier. According to Ref. [12] the interfacial potential can be estimated from an equivalent circuit model, which is depicted in Fig. 3.13. According to this model V_I reads

$$V_I(t_1) = \frac{1}{C_{tot}} \int_{t0}^{t1} I_{tun}\,(d_{eff}(x), V_I)\,dt + \frac{C_I}{C_{tot}}\left(V_{in} - V_S(x,\phi_B)\right) - \frac{1}{R_I(x)\cdot C_{tot}} \int_{t0}^{t1} V_I\,dt \qquad (3.8)$$

Here V_{in} is the external applied bias voltage, V_S is the effective voltage across the Schottky contact, R_I is the resistance of the Nb_xO_y layer, I_{tun} is the tunneling current, and C_{tot} is the total capacitance of the device containing C_T and C_I, the capacitances of the tunneling and Nb_xO_y layer, respectively. Further, x is the memristive state variable which is a measure for the charge concentration within the Nb_xO_y layer and which varies the effective tunneling distance d_{eff}, as well as R_I and V_S. In more detail, the first term of Eq. 3.8 describes the electronic conduction through the tunnel barrier, sketched by a current source in the equivalent circuit model. The second term of Eq. 3.8 comes from the voltage drop at the Schottky barrier, while the last term accounts for an interface potential dependent leakage current within the Nb_xO_y layer. By using the transport equation presented in Sec. 2 for I_{tun} and V_S, *I-V* characteristic presented in the Fig. 3.13 is obtained which nicely reflect the experimental current-voltage characteristic [12].

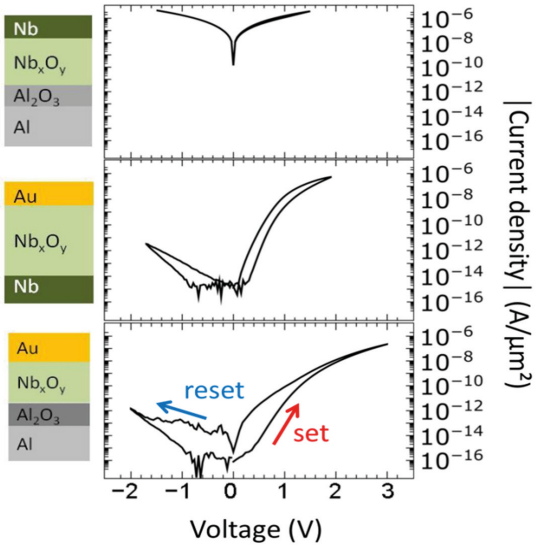


Fig. 3.12 Absolute current density |J| versus applied bias voltage of an Al/Al$_2$O$_3$/Nb$_x$O$_y$ tunnel junction, a Nb/Nb$_x$O$_y$/Au Schottky contact and, for comparison, the current-voltage characteristic already depicted in Fig. 3.11 with an Al/Al$_2$O$_3$/Nb$_x$O$_y$/Au layer sequence.On the left columnthe simplified cross-sectional view of the devices.

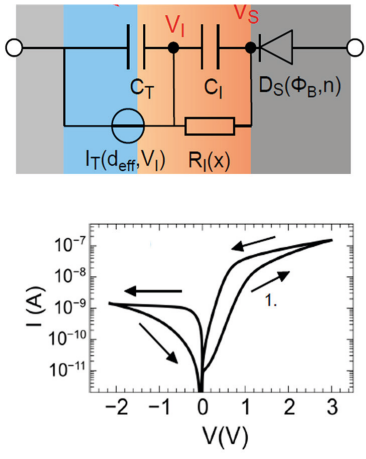


Fig. 3.13: Equivalent circuit model and calculated I-V curve. Adapted from [12]

Applications and Retention times

The most promising feature of this memristive device concept is that the resistive switching is restricted to interface effects. The use of homogenous, interface effects as the origin of memristive switching avoids the formation of active current paths (conductive filaments) through the device. This avoids the drawbacks of initial electroforming steps and poor performance reproducibility and allows the targeted development of memristive devices by a controlled modification of interfacial potentials. Further, the finite activation energies of the ionic species lead to a frozen (memory) resistance state in case of zero bias and improve therefore the data retention as shown in Fig. 3.14. Therein, the retention characteristic of the memristive double barrier device is compared to the retention characteristic of the single Schottky barrier memristive device, which shows that the introduction of the Al_2O_3 tunnel barrier led to a significantly improved retention characteristic.

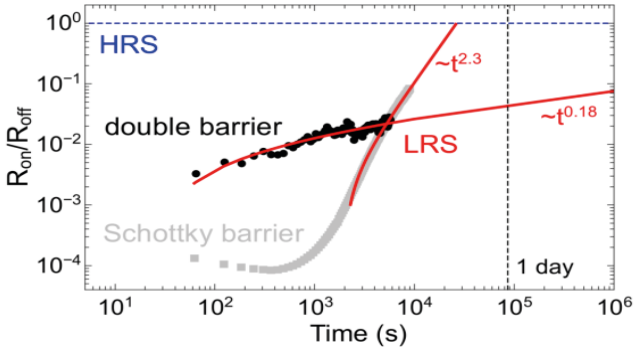


Fig. 3.14: *Comparison of the retention characteristics of a Nb_xO_y /Au Schottky contact and an double energy barrier* $Al/Al_2O_3/Nb_xO_y/Au$ *device. Adapted from [12]*

As possible applications field of the memristive tunneling devices, the field of neuromorphic computing has been proposed in Ref. [12], where high resistive devices are desired in order to reduce the overall power consumption of the whole neural system. A drawback of the relative high device resistances seems to be that the scalability of the device is therewith restricted. Here, the use of other electrode materials might be necessary to reduce the overall OFF resistance of the tunneling device.

3.3 Ferroelectric Tunneling Junctions

Ferroelectric materials possess an unique physical property: They exhibits a spontaneous *electric* polarization which can be electrically switched between two possible orientations. In 1971, Esaki et al. proposed the idea of ferroelectric tunnel junctions (FTJs), that time termed as a *"polar switch"*, in which the insulating barrier of the tunnel junction is replaced by a thin ferroelectric barrier [23]. It was anticipated that introducing a thin ferroelectric barrier could further enhance the functional properties of the tunnel junctions. For example, switching the polarization of such ferroelectric barriers was expected to modify the tunnelling probability of electrons and, hence, could have a pronounced effect on the resistance of the junction, leading to resistive switching (RS) effect. The RS effect driven by ferroelectric nature of the barrier is commonly known as the *tunneling electoresistance* (TER) effect. Figure shows the basic operation scheme of the proposed "polar switch".

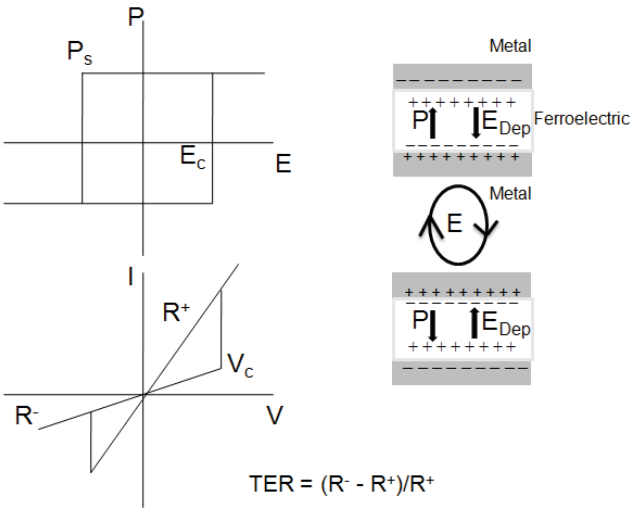


Fig. 3.15: *Schematic of the two resistive states "polar switch". Adapted from [23]*

Working Mechanisms

FTJs require ferroelectric nanolayers with switchable spontaneous polarization at a thickness of just a few unit cells. However, due to the lack of the capabilities in fabricating ultrathin ferroelectric films, the experimental realization of FTJs eluded the scientific community for almost three decades. Nevertheless, recent advancements in epitaxial thin film growth techniques and the ability to achieve ferroelectricity in ultrathin films grown on suitable substrates, attributed to the enhancement of the out-of-plane polarization by substrate-induced lattice strains and the elastic stabilization of a single-domain state, triggered high research activities to realize the FTJs [24-26].

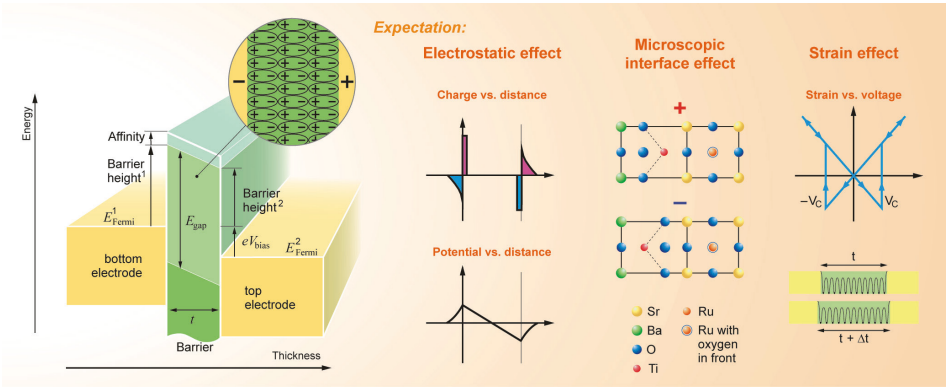


Fig. 3.16: *Schematic of the three different mechanisms affecting tunneling electroresistance in FTJs. Adapted from [28]*

A new era in the research activity of FTJs started in 2003 with the observation of resistive switching (RS) behavior in a *6* nm thick $Pb(Zr_{0.52}Ti_{0.48})O_3$ (PZT) layer integrated metal-ferroelectric-metal junctions. These results triggered intensive experimental and theoretical research in the field of FTJs [27,29,30]. As summarized in Fig. 3.16, the origin of the TER effect in FTJs is mainly attributed to three different mechanisms that could affect the tunneling probability of electrons on polarization switching by changing (a) the electrostatic potential across the tunnel junction, (b) interface bonding strength and orbital hybridization (c) strain effect associated with piezoelectric response [28,31].

Realization of FTJs

Given that the application of a high electric-field (*1 – 5* MV/cm) to the junction for tunneling electroresistance measurements, there remains the probability that electrochemical ('redox'-based) RS effects might be the cause of the nonvolatile RS, often misleading the observers [32]. Such redox-based RS effects in general encompass conducting filament formation within the insulating matrix as a consequence of dielectric breakdown due to the high electric-field. Therefore, it is of crucial importance to show the occurrence of tunneling electroresistance effect at voltages corresponding to the coercive fields of the ferroelectric tunnel barrier.

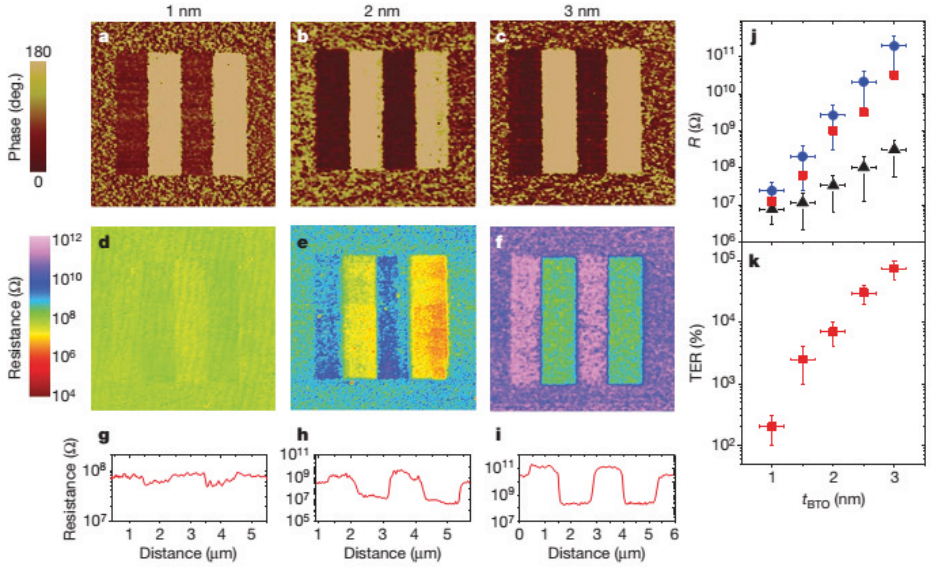


Fig. 3.17: *PFM phase image (a-c), C-AFM resistance map (d-f) and the corresponding resistance (g-i) profiles of the positively and negatively poled areas for the ultrathin $BaTiO_3$ films ranging between 1-3 nm. (j) Thickness dependence of the resistance for the positively (black triangles) and negatively poled (blue circles) domain areas. (k) Thickness dependence of the TER ratio. Adapted from [35]*

Scanning probe microscopy (SPM) techniques allow the simultaneous probing of polarization by piezoresponse force microscopy (PFM) and tunneling current by conductive atomic force microscopy (C-AFM) to distinguish between the ferroelectric driven and redox-based RS effects [33]. This approach has successfully been applied to investigate the effect of polarization

switching on the transport properties of thicker ferroelectric films normally prohibit direct tunneling [34]. The observed large modulation of the current, in the regime of electron tunneling assisted by a high electric field (Fowler-Nordheim tunneling), on ferroelectric switching was explained with the change in the Schottky barrier height formed at the ferroelectric-metal tip interface [34].

Recently, Garcia et al. successfully reported the correlation between ferroelectric switching (see Fig. 3.17) and the tunneling electroresistance (cf. 3.17(d)-(f)) for a ultrathin $BaTiO_3$ films ranging between 1 and 3 nm by combining PFM and CAFM techniques at room temperature [35]. The resistance of the $BaTiO_3$ barrier increases exponentially with varying thickness, a strong indication for the direct tunneling of electron through the barrier. In addition, a large value of TER ratio ≈ 75,000% ($TER_{ratio} = (R^- - R^+)/R^+$, where R^- and R^+ denote the FTJ's resistances corresponding to *P*+ and *P*- polarization states of the ferroelectric tunneling barrier) was reported for a 3 nm thick barrier and decreasing with decreasing the barrier thickness [35]. Recently, a lot of promising results addressing the role of ferroelectric/electrode interfaces and their influence on the magnitude of TER ratio have been reported [36,37]. Despite recent advances in experimental and theoretical studies of FTJs, many questions concerning their electrical behavior remain still open. In particular, the choice of electrode materials, role of defects, understanding the kinetics of TER, separation of the ferroelectric-driven TER effect from electrochemical ('redox'-based) resistance-switching effects and reliability related issues have to clarify in future to avail the full potential of FTJs. The recent progress in ferroelectric and multiferroic tunnel junctions tunnels are summarized in ref. [38,39].

4 Applications in Neuromorphic Systems

The application of memristive devices in neuromorphic circuits gained considerable interest in the last year's and based on the fact that the conductance of memristive devices can be precisely adjusted by changing the duration and amplitude of the applied voltage. Respect to this important synaptic functionalities have been mimicked so far [40].

In this section we like to show evidence that especially memristive tunneling devices are promising candidates for neuromorphic systems due to their continuous and homogeneous resistance switching behavior, device variability, current-voltage non-linearity, and high resistivity. In Sec. 4.1 the concept of *Hebbian* plasticity is presented, which allows to emulate some of the key cellular mechanisms of learning and memory. Thereafter in Sec. 4.2, device requirements for the emulation of *Hebbian* plasticity based on memristive devices are presented in order to show the application potentials of memristive tunneling devices.

4.1 Memristive *Hebbian* Plasticity

The basic building blocks of every neural network are neurons and their inter-cellular connections, called synapses. In nature, synapses are playing a crucial role for learning and memory, since they are plastic, which means that they change their state depending on the neural activity of the respectively coupled neurons [41]. Synaptic plasticity ensures a temporary *potentiation* or *depression* of inter-cellular connections. At the cellular level, the famous *Hebbian learning rule*, which states *"neurons that fire together wire together"*, allows to define a mathematical framework for cellular learning processes:

$$\frac{d\omega}{dt} = F(\omega(t), v_{pre}(t), v_{post}(t)) \tag{4.1}$$

Here, $\omega(t)$ is the synaptic weight, while $v_{pre}(t)$ and $v_{post}(t)$ are the activities of the pre- and post-synaptic neuron, respectively. In particular, the *Hebbian* learning rule implys that the synaptic coupling strength is affected, if both, the pre- and the post-synaptic neurons are simultaneously active. In the simplest Hebbian learning model F can expressed by $F = \beta v_{pre}(t) v_{post}(t)$. In biological systems learning and memory processes underlying three major properties*: locality*, *cooperativeness* and *associativity*, as illustrated in Fig. 4.1 which has been adapted from [42]. While *locality* means that the change of synaptic efficacy is critically depending on the activity of two interconnected neurons by a specific synapse but not on the activity of other neurons in the network, *cooperativeness* imply that the post- and pre-synaptic neuron must be simultaneously active. The third aspect, *associativity* bridges the gap from the one-dimensional cellular learning mechanism to the multi-dimensional network level. From the above it follows that neurons in networks face a competitive situation in a way that synaptic weights grow at the expense of others [41].

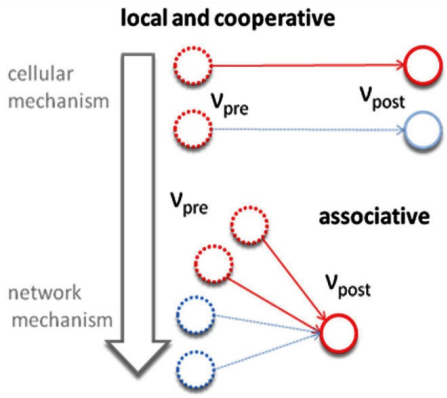


Fig. 4.1. *Illustration of the key properties of Hebbian learning, which are locality cooperativity and associativity; adapted from [42]*

The fundamental property of a memristive system is that its device resistance R or conductance G is depending on the history of the applied voltage V (cf. Eq 1.1). Regarding this property, the change of the memristive state can be used for the emulation of the synaptic weight change, such as [43]

$$\frac{d\omega}{dt} = \frac{dG(V,t)}{dt} = f(G,V,t). \tag{4.2}$$

Here, f is a continuous function describing the dynamics of the resistance switching process. Comparing Eq. 4.2 with Eq. 4.1 we can identify f as the synaptic weight change within the *Hebbian* plasticity model.

4.2 Device Requirements for the Emulation of *Hebbian* Plasticity

The wide variety of memristive materials, devices and their working conditions make the search for the *"best"* memristive device for neural systems quite complicate. In particular, this poses the question, which are the desired device requirements? In order to give an answer to this question it might be helpful to have a plasticity model at the hand, which accounts for

biological data and which is suitable to describe common plasticity measurements of memristive devices. In particular, such a model should only depend on parameters which are available from current-voltage measurements and should be applicable to different device concepts to compare them among each other. In Ref. [42] a phenomenological model is presented which fulfils these requirements. In the following this model is briefly introduced in order to discuss the desired neuromorphic device requirements.

Plasticity Model

The basic idea of the phenomenological learning model of Ref. [42] is to use the logistic differential equation to describe the synaptic weight change, since the logistic differential equation provides a strong synaptic weight change at the beginning, which gradually decreases with increasing weight and converges to zero if the maximal synaptic weight is reached. Therefore, Eq. 4.1 can be reformulated as

$$\frac{d\omega}{dt} = \beta(\omega)\,\omega(t)\left(1 - \frac{\omega(t)}{\omega_{max}}\right) \tag{4.3}$$

where $\beta(\omega)$ is a weight dependent learning rate and ω_{max} the maximal synaptic weight. To solve this equation the processes of synaptic *potentiation* and *depression* can be regarded independently:

$$\begin{aligned} \omega^P(t) &= \omega_0(t_0) + \frac{\omega_{max}}{1 + \exp(-\beta_P(t - t_0))} \\ \omega^D(t) &= \omega_0(t_0) + \frac{\omega_{max}}{1 + \exp(-\beta_D(t - t_0))} \end{aligned} \tag{4.4}$$

Here β_P and β_D are the distinct learning rates for the *potentiation* and *depression* process, respectively, and ω_0 is the inertial synaptic weight at the time t_0. Further, following Ref. [42] the learning rates β_P and β_D are conductance dependent (according to Eq. 4.1) and might be described by

$$\begin{aligned} \beta_P(V(t), G(t)) &= K_P\, G(V, t) \\ \beta_D(V(t), G(t)) &= K_D\, G(V, t) \end{aligned}\,, \tag{4.5}$$

where K_P and K_D are positive parameters describing the kinetics of the weight change process. Thus, β provides both, a weight and voltage dependence of the plasticity model [42].

For the plasticity emulation with memristive devices voltage trains are typical used which consist of a set of n equivalent positive voltage pulses (*potentiation* pulses) followed by n negative voltage pulses (*depression* pulses). Regarding this voltage scheme, the continuous time and voltage used in Eqs. 4.3 - 4.5 can be replaced by $t = n\Delta t$ and ΔV, where Δt and ΔV are the width and amplitude of an individual applied voltage pulse. The resulting *plasticity* curves are shown in Fig. 4.2. Therein the normalized weight changes for *potentiation* and *depression* at the interval $t = n\Delta t$ ($\Delta t = 1ms$) are shown. As a result, β_p is increasing during potentiation (cf. left inset in Fig. 4.2), while β_d decreases under depression condition (cf. right inset in Fig.4.2). The resulting weight evolutions are depicted with red lines. In particular, the actual weight change $\Delta\omega(t)$ depends on previous weight changes, which is at the heart of memristive systems. Thus, memristive learning rates significantly differ from constant learning rates where $\beta^{p(d)}$ is constant (compare black curves in Fig. 4.2).

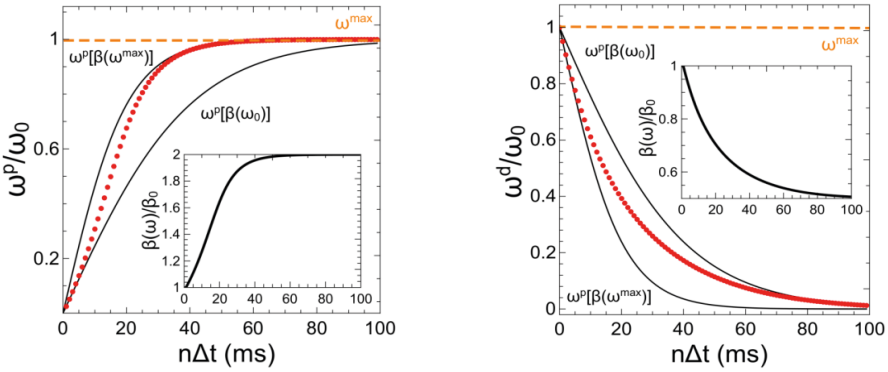


Fig. 4.2: *Normalized weight change as a function of pulse numbers n with constant pulse width Δt (1 ms) and height ΔV according to Eqs. 4.3-4.5 illustrating the weight saturation and dynamics of the learning rate. Here, black curves correspond to constant learning rates β= β(ω$_{max}$) and β= β(ω$_0$) for the maximal ω$_{max}$ and minimal weight ω$_0$, respectively. Red curves are obtained from dynamical learning rates β= β(ω) according to Eq. 4.5 assuming that β is proportional to ω(t) as explained in the text. The evolution of the particular β functions are depicted in the insets. Adapted from [42]*

Device Requirements

It follows from above that the resistance switching characteristic of the particular memristive device strongly influences the synaptic weight evolution. In particular, a continuous change in the resistance of a memristive device gradually changes the learning rate β (cf. Fig. 4.2) and it is therefore close to biological plasticity [42]. Regarding these memristive devices which exhibit gradual homogeneous switching characteristics (as it is the case for most of the memristive tunnel junctions) are preferred to memristive devices exhibiting a binary resistance switch. Furthermore, Fig. 4.2 implies that the desired device should exhibits a multiple number of resistance states combined with a high *LRS/HRS* ratio. In order estimate the desired resistance switching speed of memristive devices, it is worth to consider the biological time scale of an action potential which is in the range of *2* ms - *3* ms. Regarding this, ms duration voltage pulses should be applicable to the device in order to vary the device conductance within the biological time scale. This, in fact, makes ionic devices like the memristive tunnelling devices presented in Sec. 3.2 very attractive, since the ionic diffusion times are typically at this time scale.

It is important to mention that learning in biological system manifests in network behaviour, where synaptic plasticity is a (local) cellular precondition. Hence, depending on the particular network topology, specific requirements for the single memristive devices can differ. Furthermore, for network implementations, memristive devices have to be characterised on a wafer scale, which allows getting statistics of device parameters from a large number of devices and switching events. This should allow coming to suitable models of devices needed for circuit designers. Network implementation is so far challenging, the two main issues to be addressed being device variability (or reproducibility) and *I–V* nonlinearity [6]. Considering these points, memristive tunnelling devices seems to be promising candidates. Particularly, since for those devices the resistance switching process is restricted to interfacial processes, this ensures lower device variability and allows achieving defined *I-V* nonlinearities. As discussed in Sec. 3, the MemFlash cell device concept is very convincing due to the compatibil-

ity with Si-technology and by the opportunity of modelling the device using a simple capacitive model. In terms of the power consumption and switching times, ionic and ferroelectric tunnel junction can further convince.

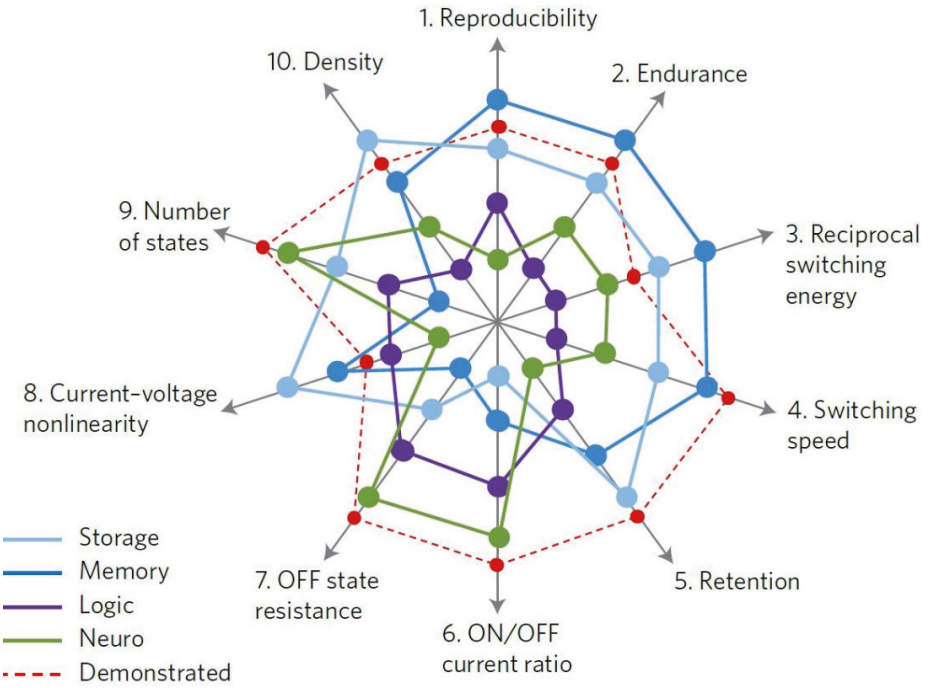


Fig. 4.3: *Device performance requirements are ranked (qualitatively) among the considered applications. A higher position on the axis implies a higher required value of the specific metric. Adapted from [6]*

References

[1] Dearnaley, G., A. M. Stoneham, and D. V. Morgan. "Electrical phenomena in amorphous oxide films." *Reports on Progress in Physics* 33.3 (1970): 1129.

[2] Strukov, Dmitri B., et al. "The missing memristor found." *Nature* 453.7191 (2008): 80-83.

[3] Chua, Leon O. "Memristor-the missing circuit element." *Circuit Theory, IEEE Transactions on* 18.5 (1971): 507-519.

[4] Ha, Sieu D., and Shriram Ramanathan. "Adaptive oxide electronics: A review." *Journal of Applied Physics* 110.7 (2011): 071101.

[5] Waser, Rainer, et al. "Redox-based resistive switching memories–nanoionic mechanisms, prospects, and challenges." *Advanced Materials* 21 (2009): 2632-2663.

[6] Yang, J. Joshua, Dmitri B. Strukov, and Duncan R. Stewart. "Memristive devices for computing." *Nature nanotechnology* 8.1 (2013): 13-24.

[7] Sze, Simon M., and Kwok K. Ng. *Physics of semiconductor devices*. John Wiley & Sons, 2006.

[8] Simmons, John G. "*Generalized formula for the electric tunnel effect between similar electrodes separated by a thin insulating film." Journal of Applied Physics* 34.6 (1963): 1793-1803.

[9] Ziegler, M., et al. "Memristive operation mode of floating gate transistors: A two-terminal MemFlash-cell." *Applied Physics Letters* 101.26 (2012): 263504.

[10] Riggert, C., et al. "MemFlash device: floating gate transistors as memristive devices for neuromorphic computing." *Semiconductor Science and Technology* 29.10 (2014): 104011-104019.

[11] Maier, P., et al. "Memristive operation mode of a site-controlled quantum dot floating gate transistor." *Applied Physics Letters* 106.20 (2015): 203501.

[12] Hansen, M., et al. "A double barrier memristive device." *Scientific reports* 5 (2015).

[13] Sawa, Akihito. "Resistive switching in transition metal oxides." *Materials today* 11.6 (2008): 28-36.

[14] Mikheev, Evgeny, et al. "Resistive switching and its suppression in Pt/Nb: SrTiO3 junctions." *Nature communications* 5 (2014).

[15] Aoki, Yoshitaka, et al. "Bulk mixed ion electron conduction in amorphous gallium oxide causes memristive behaviour." *Nature communications* 5 (2014).

[16] Baik, Seung Jae, and Koeng Su Lim. "Bipolar resistance switching driven by tunnel barrier modulation in TiOx/AlOx bilayered structure." *Appl Phys Lett* 97.7 (2010): 072109.

[17] Jeong, Doo Seok, Byung-ki Cheong, and Hermann Kohlstedt. "Pt/Ti/Al2O3/Al tunnel junctions showing electroforming-free bipolar resistive switching behavior." *Solid-State Electron.* **63**, 1 (2011).

[18] Hu, Jenny, et al. "Impact of fixed charge on metal-insulator-semiconductor barrier height reduction." *Applied Physics Letters* 99.25 (2011): 252104.

[19] Park, C., et al. "Electrode-dependent electrical properties of metal/Nb-doped SrTiO3 junctions." *Journal of Applied Physics* 103.5 (2008): 4106.

[20] Baikalov, A., et al. "Field-driven hysteretic and reversible resistive switch at the Ag-Pr0. 7Ca0. 3MnO3 interface." *Appl. Phys. Lett.* **83,** 957 (2003).

[21] Kohlstedt, H., K-H. Gundlach, and S. Kuriki. "Electric forming and telegraph noise in tunnel junctions." *Journal of applied physics* 73.5 (1993): 2564-2568.

[22] Meyer, Rene, et al. "Oxide dual-layer memory element for scalable non-volatile cross-point memory technology." *Non-Volatile Memory Technology Symposium, 2008. NVMTS 2008. 9th Annual*. IEEE, 2008.

[23] Esaki, L., R. B. Laibowitz, and P. J. Stiles. "Polar switch." *IBM Tech. Discl. Bull* 13.2161 (1971): 114.

[24] Tybell, Thomas, C. H. Ahn, and Jean-Marc Triscone. "Ferroelectricity in thin perovskite films." *Applied Physics Letters* 75.6 (1999): 856.

[25] Junquera, Javier, and Philippe Ghosez. "Critical thickness for ferroelectricity in perovskite ultrathin films." *Nature* 422.6931 (2003): 506-509.

[26] Pertsev, N. A., and H. Kohlstedt. "Elastic stabilization of a single-domain ferroelectric state in nanoscale capacitors and tunnel junctions." *Physical review letters* **98**.25 (2007): 257603.

[27] Contreras, J. Rodrıguez, et al. "Resistive switching in metal–ferroelectric–metal junctions." *Applied physics letters* 83.22 (2003): 4595-4597.

[28] Tsymbal, E. Y. & Kohlstedt, H. "Tunneling across a ferroelectric". *Science* **313**, 181 (2006).

[29] Kohlstedt, H., Pertsev, N. A. and Waser, R., *Materials Research Society Symposium C, Ferroelectric thin films X*, *Proceedings* **688,** 161-172 (2002).

[30] Zhuravlev, M., Ye Sabirianov, R. F., Sabirianov, Jaswal, S. S. and E. Y. Tsymbal, E. Y., "Giant Electroresistance in Ferroelectric Tunnel Junctions", *Phys. Rev. Lett.* **94**, 246802 (2005).

[31] Kohlstedt, H., Pertsev, N. A., Rodríguez Contreras, J., and Waser, R., "Theoretical current-voltage characteristics of ferroelectric tunnel junctions", *Phys. Rev. B* **72**, 125341 (2005).

[32] Soni, Rohit, et al. "Bipolar switching polarity reversal by electrolyte layer sequence in electrochemical metallization cells with dual-layer solid electrolytes." *Nanoscale* 5.24 (2013): 12598-12606.

[33] Kohlstedt, H., et al. "Method to distinguish ferroelectric from nonferroelectric origin in case of resistive switching in ferroelectric capacitors." *Applied Physics Letters* 92.6 (2008): 2907.

[34] Maksymovych, Peter, et al. "Polarization control of electron tunneling into ferroelectric surfaces." *Science* 324.5933 (2009): 1421-1425.

[35] Garcia, Vincent, et al. "Giant tunnel electroresistance for non-destructive readout of ferroelectric states." *Nature* 460.7251 (2009): 81-84.

[36] Soni, Rohit, et al. "Giant electrode effect on tunnelling electroresistance in ferroelectric tunnel junctions." *Nature communications* 5 (2014).

[37] Wen, Zheng, et al. "Ferroelectric-field-effect-enhanced electroresistance in metal/ferroelectric/semiconductor tunnel junctions." *Nature materials* 12.7 (2013): 617-621.

[38] Tsymbal, E.Y., Gruverman, A., Garcia, V., Bibes, M., and Barthélémy, A., "Resistive switching phenomena in thin films: Materials, devices, and applications Ferroelectric and multiferroic tunnel junctions", *MRS Bulletin* **37** 138 (2012).

[39] Garcia, V. & Bibes, M., "Ferroelectric tunnel junctions for information storage and processing", *Nature Communicatons* **5**, 4289 (2014).

[40] Jeong, Doo Seok, et al. "Towards artificial neurons and synapses: a materials point of view." *RSC Advances* 3.10 (2013): 3169-3183.

[41] Gerstner, Wulfram, and Werner M. Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[42] Ziegler, Martin, et al. "Memristive Hebbian Plasticity Model: Device Requirements for the Emulation of Hebbian Plasticity Based on Memristive Devices." (2015).

[43] Zamarreño-Ramos, Carlos, et al. "On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex." *Frontiers in neuroscience* 5 (2011).

E1 Reliability of Memristive Elements

Arne Heittmann[1], Tobias G. Noll[1], Dirk J. Wouters[2], Yang-Yin Chen[2]
Andrea Fantini[2], Nagarajan Raghavan[2,5]

[1] RWTH Aachen, EECS, Schinkelstr.2, D-52062 Aachen, Germany

[2] imec, Kapeldreef 75, B-3001 Leuven, BELGIUM; now: IWE 2, RWTH Aachen, Germany

[3] also with KULeuven, ESAT, Arenberg Park 10 B-3001 Leuven, Belgium

[4] currently with RWTH Aachen, IWE II, Somerfeldstraße 24, D-52074 Aachen, Germany

[5] currently with Singapore University of Technology & Design (SUTD), Singapore – 138 682.

Contents

1 Introduction **2**
1.1 Non-Volatile Memories 3

2 Reliability of Bipolar Switching TMO RRAM **3**
2.1 Endurance 4
2.2 Retention 7
2.3 Variability 9
2.4 Random Telegraph Noise (RTN) 13
2.5 Read-Disturb and Write-Disturb 18
2.6 Summary 20

3 Monte-Carlo Circuit Simulation of RRAM Circuits **20**
3.1 ECM Variability Model for Circuit Simulation 20
3.2 Simulation of Write-Modify Cycles in Passive Crossbars 23
3.3 Summary 29

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

In modern integrated systems circuit designers are faced with a multi-objective optimization problem. An "optimal" circuit would satisfy conditions of:

- Minimal circuit area
- Minimal power consumption
- Maximum performance
- Maximum noise immunity
- Maximum yield

It can be shown that these objectives are in fact competing. For instance, in CMOS integrated circuits high speed circuits comprise transistors with low threshold voltages which inevitably let the static power consumption rise. Architectures featuring units capable of processing information in parallel offer a significant speed-up in performance at the cost of area for implementing these units in parallel at the physical level. Parametric yield improvement can be realized in many cases by providing larger margins for power supply and signal amplitudes which directly results in a worse power consumption for the system.

Following the trend of device miniaturization down to the nanoscale reliability becomes a major concern for the design of circuits. In the same way as the device count (e.g. transistors, memory elements, etc.) in an integrated system rises in order to enable more and more complex functions, parametric variability of physical device parameters gets enlarged. As a consequence, putting arbitrary (or at least very beneficial) margins for performance, power supply, etc. into the specifications is no longer possible. If the impact of parametric variability is overestimated, typically an integrated system features

- Larger chip area (e.g. by placing redundancy circuits which are not needed)
- Larger power consumption (e.g. by scaling up the supply voltage)
- Suboptimal circuits (e.g. by rejecting of good design options)
- Larger design time (e.g. meeting the design specification becomes harder by assuming an unrealistic worst case condition)

Overestimation of variability essentially increases the design effort.

On the other hand, if the impact of variability is underestimated, the consequences are

- Yield loss (e.g. additional redundant circuits would be necessary)
- Performance reduction (e.g. by missing the "realistic" critical path)
- Complicated debugging (e.g. in the absence of a realistic variability-aware simulation model)

Underestimation of variability sets the burden to the manufacturing effort in order to obtain a reliable circuit which meets the original performance requirements. This option, however, becomes complicated in the future. Last but not least as both, the presence or the absence of individual particles at the atomic scale can have significant parametric impact on the device performance. These factors are hard to control during fabrication. Other sources of variability are of pure statistical nature, such as the random dopant deposition in a MOSFET's channel region or the trapping and de-trapping of charges observed as random telegraph noise (RTN).

1.1 Non-Volatile Memories

Apparently, nanoscale circuits need to be tailored with respect to device variability. Here, realistic models that express the impact of variability are required. This is especially true for memory technologies where memory cells are regarded as High-Replication Circuits (HRC) demanding for an extremely high parametric yield. Besides speed, power consumption, and area requirement memories have additional characteristic features to fulfill for the entire life cycle. For non-volatile memories the following features are of major interest:

- Maximum endurance
- Maximum retention

Endurance is quantitatively defined as the minimum number of Program/Erase cycles an individual cell is able to withstand without showing faulty behavior. In the worst case the exposed stress results in a cell defect, i.e. the cell is not functional anymore. In addition to a stress-induced cell defect it is necessary to have the possible cell states (which represent the stored information) suitably separated in order to be able to sense the cell's state correctly.

Retention describes the ability of a memory cell to keep the stored information over an extended period of time (e.g. 10 years). There are several factors which have influence on the retention. First, there is temperature which contributes to almost any temperature-activated ageing process in integrated circuits. Therefore, along with retention a temperature budget has to be defined. Second, architectural features of a cell array contribute to an accelerated degradation of a cell's state. Typically, memory cells are arranged with maximum density which implies that word lines and bit lines used to access a cell are connected to several cells. When a new state is written to a cell not only the accessed cell is getting exposed to a "write stress", also neighboring cells are exposed (at least to a fraction of) this stress which potentially leads to a disturbance of their state (write disturb). Also a read access can result in a small disturbance. Although this disturbance is small it may accumulate to a significant disturbance if the number of read accesses gets large (read disturb).

The drive for the development of new emerging memory technologies as resistive switching RAM, is coming from scaling issues that most -if not all- of the current (charge based) memory technology face beyond the 1Xnm technology node. While we do see a steady improvement of the reliability characteristics of RRAM as the technology is becoming better controlled, it is important to assess what are the real limitations (and/or possible mitigations) to assess the potential of RRAM to substitute for one of these technologies in further scaled nodes.

This chapter is divided into two parts. In the first part a comprehensive overview about the physical grounds causing reliability issues as well as experimental results are discussed. In the second part a circuit simulation model is presented which is used in Monte-Carlo simulations to predict the impact of variability on the circuit level. As an example of application for this model, a particular Write-Verify algorithm is explored and characterized.

2 Reliability of Bipolar Switching TMO RRAM

This section tries to give a comprehensive overview of the different aspects of the RRAM reliability focusing not only on "best" achieved specifications, but also on the models and understanding of the reliability physics involved. The drive for the development of new emerging memory technologies as resistive switching RAM, is coming from scaling issues

that most -if not all- of the current (charge based) memory technology face beyond the 1Xnm technology node. While we do see a steady improvement of the reliability characteristics of RRAM as the technology is becoming better controlled, it is important to assess what are the real limitations (and/or possible mitigations) to assess the potential of RRAM to substitute for one of these technologies in further scaled nodes.

2.1 Endurance

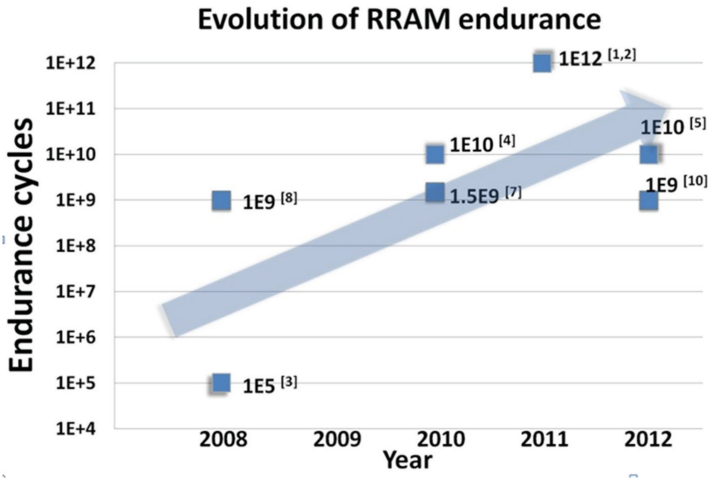


Fig. 1: *High endurance RRAM devices in publications, 2008 ~ 2012*

Most memory applications require the ability to re-write stored data. This is measured as the maximum numbers of alternative state (0/1) program/read cycles that can be applied without failure.

Filamentary- based bipolar switching transition metal oxide (TMO) RRAM in general demonstrates good cyclability. Tab.1 summarizes the achieved number of cycles for different TMO stacks, on single cell basis [1-10]. Compared to floating gate memories, which typically fail after 10^4 ~ 10^5 cycles [60], much higher endurance (up to 10^{12} cycles) is achieved endurance in RRAM. Over the years (2008 ~ 2012), the maximum endurance cycle number on RRAM also increases (Fig.1), [1-5,7-8,10]).

Table 1: *Summary of high endurance RRAM devices*

	TMOs	SET	RESET	Cycles
SAIT[1,2]	TaO based	4.5V, 10ns	7V, 10ns	10^{12}
ITRI[3,4]	HfO_2 / Ti	3.2V, 40ns	2.7V, 40ns	10^{10}
IMEC[5,6]	HfO_2 / Hf, Ti	1.8V, 5ns	1.8V, 10ns	10^{10}
HP[7]	TaO based	1.9V, 1us	2.2V, 1us	1.5 x 10^{10}
Panasonic[8,9]	TaO based	1.5V, 100ns	2V, 100ns	10^9
SEMATECH[10]	HfO_x	1.5V, 50ns	1.5V, 50ns	> 10^9

Although TMO RRAM demonstrates potential high endurance, concerns remain especially due to its defect based filamentary switching nature. Uncontrolled, stochastic changes during the defect switching may lead to variable endurance degradation, which needs to be better understood.

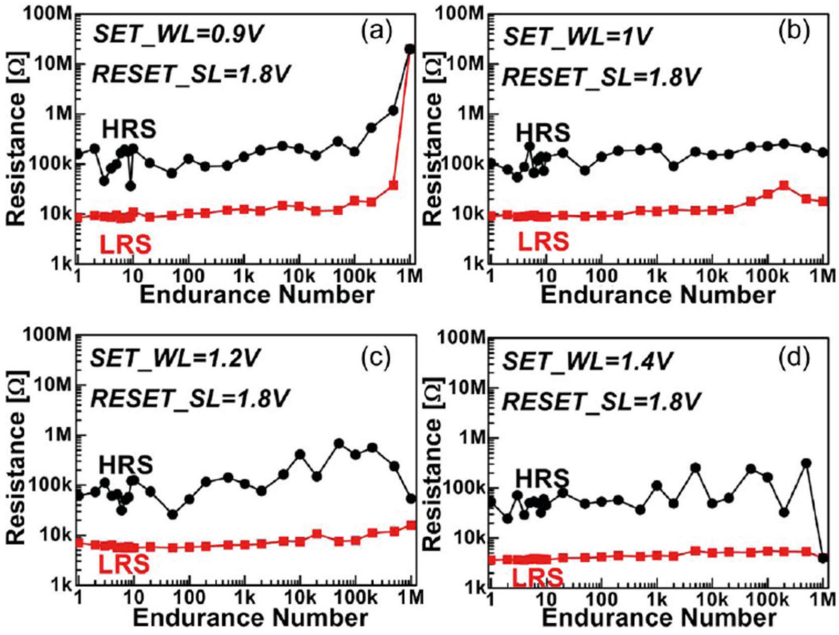


Fig. 2: *Pulse endurance behavior of the 40nm Hf / HfO2 1T1R devices, with fixed RESET pulse at WL = 3V, SL = 1.8V, 10ns. The SET pulse amplitude was varied using different WL pulses: (a) 0.9V (b) 1.0V (c) 1.2V (d) 1.4V. The SET pulse width is fixed as 100ns, 1.8V on BL. With increasing SET WL voltage, endurance failure mode shifts from LRS failure (a) to HRS failure (d). © 2012 IEEE. Reprinted, with permission, from [5].*

Based on the understanding of bipolar switching in filamentary RRAM, SET / RESET switching is achieved by drift of defects (oxygen vacancy) in the oxide. In this view, it is important to drift an equal amount of oxygen vacancies forth and back during each SET / RESET operation in order to maintain the same levels of both the low resistance state (LRS) and high resistance state (HRS). Over-SET or over-RESET, which drifts an excessive amount of oxygen vacancies towards either LRS or HRS states, will result in the possible failure of the following SET / RESET operations.

In order to balance the defect drift during SET / RESET operations, a delicate tuning of SET / RESET switching conditions is necessary [5]. Through such tuning of SET / RESET conditions, as shown in Fig.2, the failure of either over-SET or over-RESET can be avoided and endurance can be substantially improved. Though this balance point of SET / RESET operations may be material stack specific, balancing the switching operations is a general requirement for the bipolar oxygen vacancy based RRAM devices.

As demonstrated in [11-13], endurance degradation in bipolar oxygen vacancy RRAM can show up as both LRS and/or HRS state degradation. Different models are proposed to explain the degradation behaviour, which can be mainly divided into two types: failure to RESET to high enough HRS state and failure to SET to low enough LRS state. Hereby, RESET failure (HRS degradation) is typically attributed to the exhaust of oxygen (e.g. due to non-ideal drift of oxygen), resulting in incomplete recombination of oxygen and oxygen vacancies so that a larger, oxygen vacancy controlled conduction path remains. Alternatively, the filament may gradually become too big during SET (e.g. by extra oxygen generation during switching), eventually prohibiting sufficient RESET. E.g., three different RESET failure types are discussed by B.Chen et al.,[11], each with a different signature in their HRS and LRS behavior.

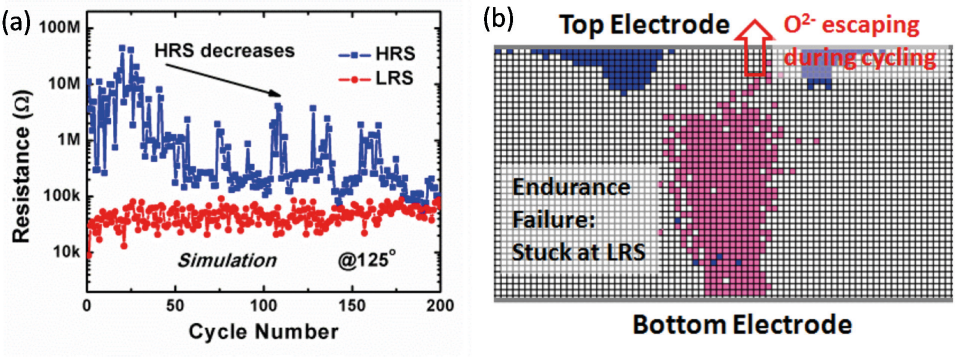


Fig. 3: *(a) Simulated endurance at 125°C. HRS decreases, and the final failure is at LRS.* ***(b)*** *V_{ox} and O^{2-} distribution at the endurance failure at the end of cycling in* ***(a)****. Insufficient O^{2-} at the interface makes the reset impossible for the RRAM cell. © 2012 IEEE. Reprinted, with permission, from [13]*

SET failure (LRS degradation), on the other hand, can be attributed to local material changes. For instance, the SET failure model in [12] is based on an increase of the SET voltage due to a local modification ("recrystallization") of the oxide material in the filament region, caused by the high temperatures (~ 895K for a 30kΩ LRS state) generated by the current during switching. As a consequence, the mobility of oxygen vacancy is reduced, and the SET operation becomes unsuccessful. In general, the switching power/energy and the temperature in the oxygen vacancy filament are important parameters impacting both LRS and HRS degradations. Due to filamentary conduction, the temperature and power distribution is strongly non uniform, an local high temperature and high power may modify or even damage the original atomic configuration of the oxide host material.

Cycling induced degradation can be further related to the thermodynamic (in)stability of the filament and the host matrix, and has to be considered in the material design of the resistive switching system. For example, Ta oxide system has exhibited a small variance from switching with cycling and thus a high endurance. The reason for this may be that this oxide system has only two stable material phases, i.e., Ta-O solid solution (filament) and Ta pentoxide (host matrix). These two material phases do not react with each other thermally to form another material phase even at high temperatures induced by Joule heating. Furthermore, there is a large oxygen solubility in the Ta-O solid solution phase, which allows for the filament to

accommodate and release oxygen ions without a phase change during cycling. An almost identical system is Hf oxide system, which has exhibited similar cycling-ability.

Simulations may help to model and better understand the endurance degradation process. E.g., a Monte-Carlo simulation on atomic scale is reported by S.Yu et al, [13]. They could simulate HRS failure, by taking into account the oxygen and oxygen vacancy recombination during switching (cf. Fig.3). The escape of oxygen during switching results in an incomplete recombination of the oxygen vacancies during RESET and leads to the HRS degradation. This again points out the importance of balancing the oxygen drift during SET / RESET conditions for a better endurance performance of RRAM devices.

2.2 Retention

Typical non-volatile memory requires 5 ~ 10 years data retention up to 85~125°C.. Various studies [6,12,14-18] have analyzed the retention behaviour of the LRS and HRS states of oxygen vacancy filamentary RRAMs. In general, after a certain period, the current (resistance) of LRS tends to decrease (increase), which indicates a degradation of the conducting filament due to a loss of oxygen vacancies. The current (resistance) of HRS shows mixed behaviours, and both increase and decrease have been observed. Fig.4 shows the typical LRS resistance increase for HfO_2 based RRAM. Understanding and optimizing the retention property of RRAM, is critical and challenging, due to the complex and even stochastic behaviour of the oxygen vacancy defects. In generally, LRS degradation is the more critical reliability issue, and will be the focus here.

Though the oxygen vacancy filamentary switching has stochastic nature, the (LRS) retention degradation on a large statistical basis follows the classical Arrhenius law dependence with temperature. Fig.5 demonstrates the Arrhenius behaviour of a typical HfO_2 / metal cap (Hf) RRAM system [12]. The LRS retention degradation is clearly accelerated with increasing stress temperature, with an extracted energy barrier E_a of 1.2 ~ 1.5 eV. The diffusion of oxygen and oxygen vacancies inside the host oxide of the RRAM cells is responsible for the retention degradation, as will be discussed below.

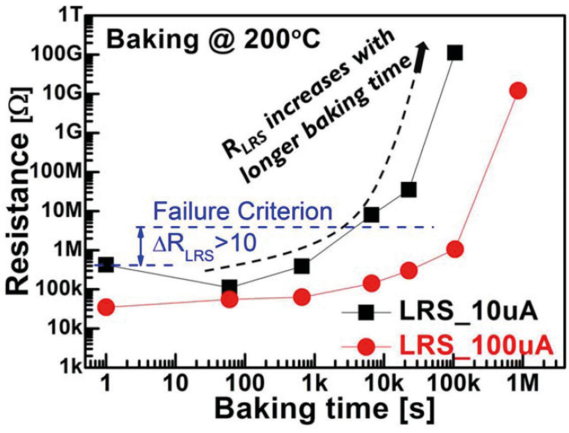


Fig. 4: *Typical LRS retention failure of a 40nm HfO_2 / Hf RRAM, for LRS programmed by either 100µA CC or 10µA CC. The LRS retention failure criterion was defined as a 10x LRS resistance increases. © 2012 IEEE. Reprinted, with permission, from [12].*

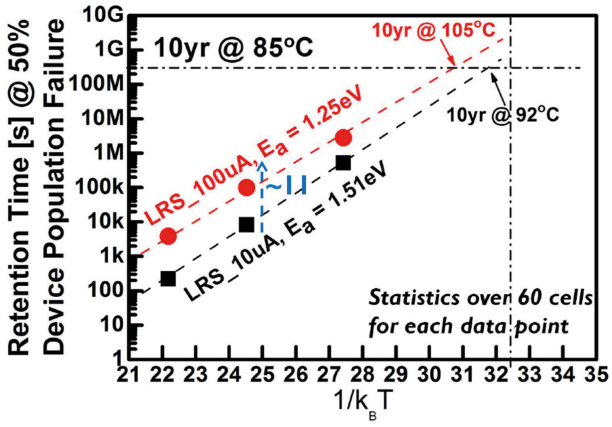


Fig. 5: *The LRS retention (programmed by 100μA and 10μA) measured from 150oC, 200oC and 250oC. The retention follows the Arrhenius law, with Ea = 1.2 ~ 1.5 eV. © 2012 IEEE. Reprinted, with permission, from [12].*

In the oxygen vacancy based RRAM, the possible (LRS) degradation mechanisms considered are (cf. Fig.6) [17]):

- Diffusion of oxygen from outside of the filament (i.e., from electrode or from surrounding oxide region) and recombination with an oxygen vacancy in the filament.
- Diffusion of an oxygen vacancy out of the filament.

Both mechanisms will result in a decrease of the number of oxygen vacancies in the filament, which leads to an increase of the LRS resistance (cf. Fig.4). Which mechanism dominates, oxygen diffusion (process 1) or vacancy diffusion (process 2), may depend on both the material stack and LRS program conditions.

A general trend on the LRS retention behaviour is that the retention improves with higher LRS current (lower LRS resistance). For lower LRS resistances, the filament consists out of more oxygen vacancies, making it more retention robust as the relative change of the filament current is smaller for the same amount of vacancies lost. I.e., the relationship between number of oxygen vacancies (n_c) in the filament constriction and the LRS resistance can be calculated using the QPC filament conduction model [15]. The non-linear relationship between number of oxygen vacancies and the resistance explains the fact that a disappearing oxygen vacancy has a much larger relative impact for a narrow filament. The current dependence of the LRS retention needs to be considered when evaluating the retention property of a particular device.

Another observation is a degradation of the data retention after cycling (at the same programming current level) [19]. The physical mechanism proposed is a loss of oxygen vacancies in the filament after cycling. This post-cycling retention is a coupled reliability issue of both cycling and retention, which depends on the cycling programming condition as well. By optimizing the cycling properties, the post-cycling retention can be improved [6].

As both the endurance and retention reliability depend on the oxygen vacancy filament, optimization of one property may impact on the other. As identified in [6], an endurance and retention trade-off can be observed in HfO_2 based RRAM, in this case by modifying the oxygen scavenging layers in the memory element stack. Oxygen scavenging layers are used in some types of RRAM devices with the effect of creating oxygen vacancies in the metal oxide, lowering the required forming conditions to create the switching filament. Using a strong oxygen scavenging layer capability, better resistance window and longer endurance is achieved. The better endurance results from the fact that the RESET failure can be compensated due to the

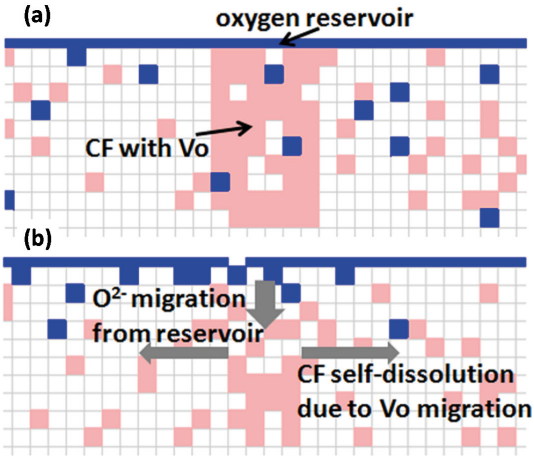


Fig. 6: ***(a)*** *An example of the initial filament configuration in LRS: pink sites are V_{ox}, blue sites are O^{2-}, the top boundary is the oxygen reservoir at the electrode/oxide interface.* ***(b)*** *An illustration of oxygen and oxygen vacancy migration processes during baking. Reproduced with permission from [17]. Copyright 2012, AIP Publishing LLC.*

larger amount of available oxygen. On the other hand, better retention can be achieved in the weak scavenging layer property case. The smaller amount of scavenged oxygen gives better retention, due to the reduced availability of oxygen that may diffuse and recombine with the oxygen vacancies in the filament.

Endurance and retention trade-off may be achieved by modifying either material or electrical parameters in the oxygen vacancy filament RRAM. Essentially, however, due to the two terminal structure of RRAM, cycling and the retention are always coupled, and have to be optimized together.

2.3 Variability

While aggressive scalability and easy manufacturability of RRAM has been demonstrated [3,20-21], one of the major issues of RRAM is the stochastic variability in the device operation that still needs further understanding to enable its commercialization.

In the well-established Flash NAND technology, random device-to-device (D2D) fluctuations of the programmed state occur as a consequence of process related device scaling. In this case we properly speak of *variability* of device characteristics like channel width, length (W, L), drain current I_{DS} and threshold voltage V_T. However, if we then look into one of these devices and we monitor the evolution of the programmed state during the device lifetime, we observe an almost constant state value between consecutive readouts (not accounting here for a long-term change of the state due to degradation mechanisms).

Resistive memories, due to their fundamentally different operating mechanism, additionally display significant intra-device, cycle-to-cycle (C2C) dispersion from the very beginning of device lifetime. In this sense we should distinguish the typical *process variability* from the *stochastic variability* where each characteristic is subject to *independent and (ideally) uncorrelated* C2C fluctuations.

As resistive memories store information as a logical value associated to the device resistance state, and transitions between these resistance states are obtained by voltage stimuli of appropriate magnitude (and polarity), the stochasticity of the RRAM operation mechanism is reflected

in both the variability of the required switching voltage, and in the variability of the resultant resistance state. Furthermore, the initial device forming step strongly influences its variability.

Forming Operation

Differently from FLASH technology, resistive memory involves a one-time initialization step called electroforming. During this step a conductive channel is created within the oxide matrix which hosts the subsequent programming operation. This forming process has been studied within the context of classic oxide breakdown theory via ramped voltage stress (RVS) or constant voltage stress (CVS) [23]. Distributions of forming voltage and its respective forming time have been shown to obey Weibull (or weakest-link) statistics for both polycrystalline and amorphous oxides. It has been shown [24] that the power involved in this operation typically affects subsequent switching operation. For instance, forming performed at much higher power than used in following switching cycles generally decreases the device resistance window and uniformity. For this reason, optimization of the forming operation is a subject of growing interest. Measures that minimize the randomness of filament formation, for example by modifying device geometry, introducing filament precursors [25] or optimizing electrical operation protocols, could help to reduce the formation variance from device to device.

Even assuming an ideal forming operation, program operations are not self-limiting as in the Flash scenario. For instance, applying voltage programming the SET operation requires a current limiting device to prevent device breakdown while, conversely, in a current driven programming, RESET operation may need a voltage limiting device to prevent degradation. In real applications the memory device, sometimes called “1R”, is thus usually coupled together with some limiting device, typically a transistor, a diode or a limiting resistor, giving rise to the so-called “1T1R”, “1S1R”, “1R1R” configurations. However it is important to realize that also these current limiting devices are themselves affected by a spread of characteristics or are subject to parasitic components like stray capacitances, the latter of these being responsible for the poor control of switching properties in the early stage of technology development.

Care has thus to be taken to decouple the characteristics of the current limiting device from those of the memory element. In this sense it is proper to speak of *intrinsic variability* of the memory device (related to the randomness of the switching process) and *extrinsic variability* associated to the combined effect of memory device and cell (and array) architecture.

From an experimental point of view, the study of the intrinsic memory device variability can be accomplished either by using nearly ideal limiting device (such as a large area long channel MOSFET) or by carefully compensating for additional contributions.

System level studies of the combined impact of the variability of both the memory and the limiting devices are however lacking, while in particular important for assessing the performance in scaled technologies at the array level. Architectural studies presented so far [27] typical focus on the impact of access line resistance, selector nonlinearity and biasing scheme on the overall power consumption assuming uniform memory device characteristics, and the power penalty introduced by non-uniformity in the limiting device itself has still to be carefully addressed.

In the early stage of RRAM development, studying unipolar switching devices, variability was mainly viewed as the spread of SET and RESET switching voltage, while variability of the programmed resistance state itself was less of a concern. This can be understood considering that unipolar operation, involving a thermally assisted rupture of the filament, is characterized by relatively high operating current and high ON-OFF ratio. As so, the resistance window did not pose large concerns, however, the RESET voltage required to disrupt the filament and the SET voltage required to reform it are subject to wide fluctuations. More importantly,

since in this case both operations are done in the same voltage polarity, an overlapping between the two voltage distributions may induce unwanted parasitic SET during RESET operation leading to strong switching instability. One approach to qualitatively model variability in this approach is linked to the "Random Circuit Breaker" model [28]. More recently, the same model was updated to take into account interfacial properties in order to model bipolar switching behaviour as well [29].

As the research focused more and more on bipolar switching devices, the need for an accurate estimation of resistance variability became prominent. This is both due to the fact that the resistance window is much smaller than in the unipolar case and to the fact that the opposite polarity of the operation voltage solves the problem of SET-RESET instability typical of unipolar devices.

Experimentally, independently from the particular oxide or stack used, all bipolar switching (oxide) RRAMs show similar switching characteristics. The resistance distribution for both ON and OFF state can in first approximation be described by means of a lognormal distribution, and to establish a useful figure of merit it is possible to consider the standard deviation of resistance distribution σ_R normalized by median resistance σ_R/R. Using this criterion it becomes apparent that by lowering the operating current not only the absolute spread increases (as it would be expected from a fluctuation proportional to filament "radius") but also the relative spread increases. In a simplified view, this behaviour is consistent with a picture where the same fluctuation of the filament radius affects much more strongly a thinner ("weaker") filament than a thicker ("stronger") filament up to the limit where the filament size may be so small that its geometry affects the conduction much more than its radius. This is captured by the σ_R/R trend in Fig.7 clearly depicting a change in regime for extremely resistive (="thin") filaments. Trying to better define what defines the filament "radius" in bipolar devices, it has been recently evidenced that switching operation involves a change into a number of conductive defects (where V_O^{2+} oxygen vacancies are generally accepted as the dominant defect in TMO oxides) either by forming a percolation path [30], a conductive subband [31], or a quantum point defined constriction [32]**.** Fig.8 illustrates the microscopic origin of resistive variability in a quantum point conduction path model [32]. A quantitative explanation for LRS resistance variability is given in [36].

In bipolar switching device the modelling and physical description of variability becomes thus linked to the role of the oxygen vacancies and consequently cannot be decoupled from the particular switching mechanism proposed. Two mainstream views are proposed: in one switching is described as generation and recombination of oxygen defects [30,31], in the other the switching action is instead caused from a (ionic) movement of the same defect without any change on their number [32,33].

- When switching is explained in terms of generation/recombination of oxygen vacancies [30], the resistance in the LRS state is produced by the creation of a percolating network of defects (oxygen vacancies) while the HRS state is due to the creation of defect-free gap due to defect recombination (i.e., oxidation of the vacancies). In this sense the spread in HRS can be justified by the random generation of defects in the gap area, while for the percolation path regeneration during SET action can be described in a similar way as standard TDDB degradation in oxides, and should follow Weibull statistics.
- When an ion-movement model is considered [34-36], ions are moving from pre-existing reservoirs located at TE and BE to enlarge or shrink (and maybe even disrupt) the filament. The source of variation in this case is linked to the number of discrete defects defining the radius of a single filament. Concerning LRS variability this view matches well

with the experimentally evidenced poissonian behaviour of resistance when modeling defects as independently emitted in time or space. In this sense the SET action describes the "effort" necessary to "nucleate" a filament either by injecting ions into a gap either by gradually changing the shape of quantum defined conducting filament.

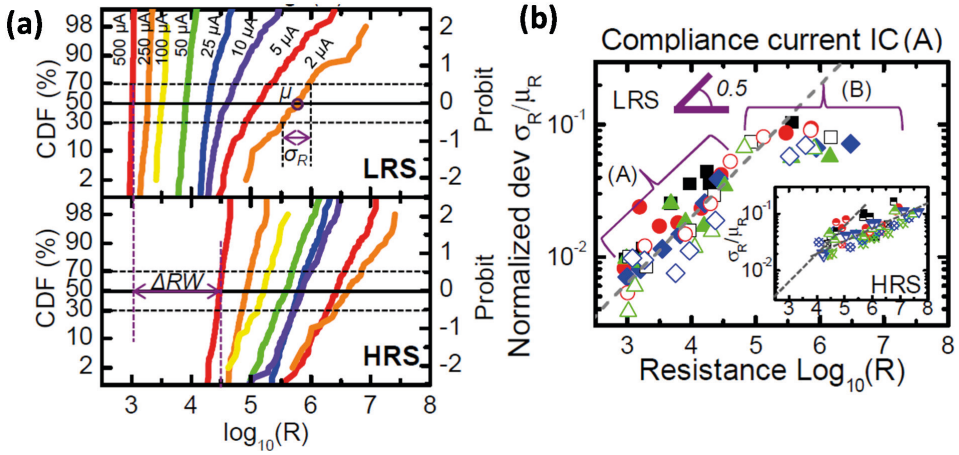


Fig. 7: ***(a)*** *Resistance distribution of LRS and HRS states obtained at different compliance currents for a TiN/HfO2/Hf/TiN stack. The decrease of operating current induces an increase of median and dispersion of resistances for both states.* ***(b)*** *Normalized standard deviation versus median resistance for LRS states and (inset) HRS states for different statistic, compliance currents and stack types. Two clearly different regions can be identified. © 2012 IEEE. Reprinted, with permission, from [35].*

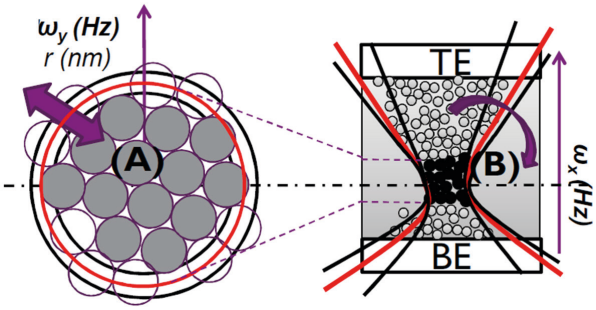


Fig. 8: *Microscopic origin of resistance variability as induced by fluctuation in the number and geometry of discrete defects defining a quantum point defined conductive path. © 2012 IEEE. Reprinted, with permission, from [35].*

2.4 Random Telegraph Noise (RTN)

Another important metric to be considered in assessing the robustness of RRAM is the random telegraph noise (RTN) phenomenon which is intrinsic to any dielectric with defects (traps). The presence of RTN can cause a large spread in the distribution of the high and low resistance state (HRS, LRS) and induce "soft errors" in reading the wrong memory state (if the memory window is relatively small). Although RTN, which is relevant mainly at read voltage (V_{READ} ~ 0.1V) conditions, does not cause irreversible damage unlike endurance test conditions, it affects the variability of the resistance distribution [37-39] and introduces large magnitude of noise in addition to the standard 1/f flicker noise and thermal white noise. Detection of RTN signals also serves as a spectroscopy tool enabling us to determine the trap properties (spatial location from dielectric – electrode interface and energy depth below conduction band of dielectric) [40]. As will be shown below, the RTN signal can also help to characterize the shape and size of the filament in the HRS state, taking the quantum point contact (QPC) formulation for the defect cluster [24].

There are two major types of RTN signals observed. One is the steady-state fluctuations involving stochastic electron capture and emission events [37,40-42] through the defects with their corresponding time constants (which depend on applied voltage, trap position and trap energy). The other components are the non-steady-state fluctuations that arise due to structural disturbances in the conducting filament (CF) due to removal or addition of oxygen vacancies [42]. It is important to note that resistive switching can occur due to two mechanisms → (1) OXRAM where switching is caused by oxygen vacancy / ion generation – recombination and drift / diffusion and (2) CBRAM (conducting bridge RAM), where switching is caused by nucleation and rupture of metallic filament due to ionic migration / electromigration.

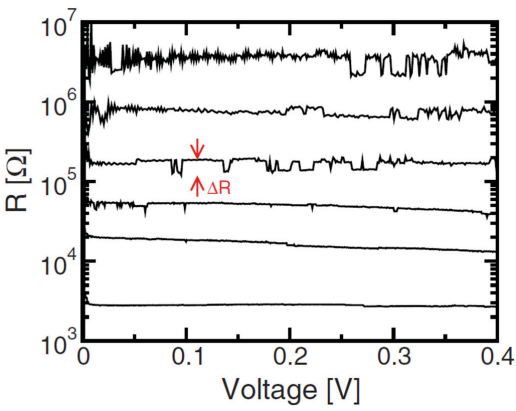


Fig. 9: *Pattern of the resistance (R = V/I) signal as a function of the resistance state. In the HRS with few defects, the RTN Lorentzian signal from every defect is more apparent, resulting in large values of (ΔI/I). In the LRS, due to a large number of defects with widely distributed time constants, the Lorentzian RTN signals add up and the fluctuations average out to give a 1/f noise trend with very low (ΔI/I). Note that the fluctuations appear to be small in the HRS as well, which is an artifact – this is because the resistance data is plotted on a logarithmic scale. Reproduced with permission from [46]. Copyright 2010, AIP Publishing LLC.*

We shall first investigate the steady-state RTN, followed by the non-steady-state component of RTN. The dependence of RTN on the compliance level, dielectric material and microstructure will also be addressed as we go along.

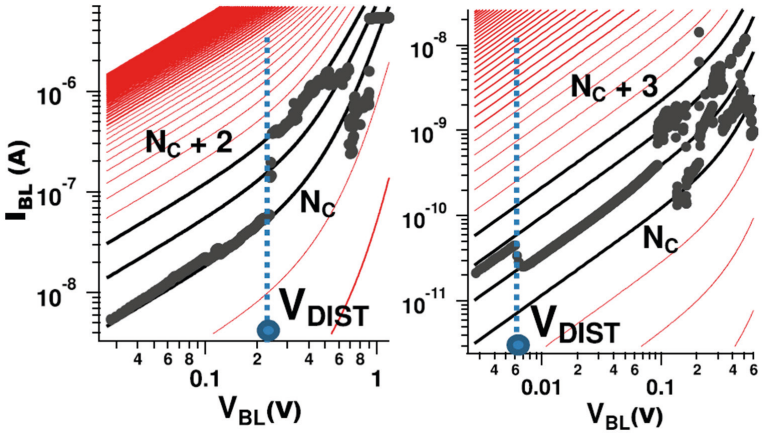


Fig. 10: *Methodology to extract the value of the disturb voltage (V_{DIST}) for every SET cycle. The I-V plots simulated for different number of constriction defects (N_C) is superimposed on to the electrical I-V measurement data. The first voltage level at which the current consistently jumps to a neighbouring I-V curve is classified as the V_{DIST} value. Note that RTN trends still exist even for $V_{BL} < V_{DIST}$, (not visible here in the logarithmic scale) however, they are small in magnitude and correspond to the electron capture – emission process. © 2013 IEEE. Reprinted, with permission, from [51].*

Considering that the fundamental physics of switching may be completely different for these two cases, it is obvious that the kinetics of RTN will also be very different. Our analysis here is predominantly focused only on RTN in OXRAM, as there are very few studies that have been carried out for RTN in CBRAM [43].

As mentioned above, the steady-state component of RTN is mainly attributed to electron capture and emission process through the oxygen vacancy defects (traps) in the dielectric, by an inelastic multi-phonon trap-assisted tunneling (ITAT) process [44]. Another possibility is the coulomb repulsion effect [45] where charged defects in the vicinity of the filament can restrict the conductivity of the filament (reduced effective filament cross-section) due to coulombic interactions. The RTN signal that is measured is an indicator of the number of "critical traps" that affect the stability of the resistance state. In the HRS, the RTN trends are more clearly observed. If the signal is sensed for a long period of time and N distinct current levels are observed (this can be detected using a Hidden Markov Model (time-lag plot) based analysis [44]), this implies that there are $log_2(N)$ traps present in the dielectric. Most signals observed are multi-level RTN where each two-level deconvoluted RTN arises from the stochastic electron capture – emission events in one single oxygen vacancy defect.

With deeper reset, there are less number of active defects and RTN signals tend to be more discrete with larger spread in the current ($\Delta I/I$) (Fig.9), as clearly shown by the statistical RTN study by Veksler *et. al.* [45] and Ielmini *et. al.* [46]. Ideally, if the RRAM can be reset to very deep states with zero defects, then we can achieve very good noise immunity. However, this is in most cases not feasible. As for the LRS state, for typical high compliance levels of 100 μA – 1 mA, there are far too many traps and the sum of many RTN signals with different time constant and current step distributions will average out to produce a *1/f* random noise signal

[47], with very low ($\Delta I/I$). Therefore, RTN in the LRS state is generally not a critical issue, unless we talk about ultra-low power switching devices with forming / SET compliance levels as low as 1 μA (which currently show very low endurance and retention).

Even in the HRS state, the values of ($\Delta I/I$) are relatively small (~10-100%) [38-39] for the charge carrier transport when compared to the vacancy fluctuations in the filament (structural disturbances), for which the value of ($\Delta I/I$) can be even an order of magnitude at times [42]. Moreover, since structural disturbances are mostly irreversible (only occasionally the current levels jump back to their initial state), their role is more detrimental in the stability of the memory state. In the next sub-section we will focus on these vacancy-induced RTN effects. Before we discuss that, some of the recent noteworthy references for carrier-induced RTN are the work done by *Lee et. al.* on TiO_x (20 nm) [40], *Puglisi et.al.* on HfO_2 (5 nm) [44] and *Ielmini et. al.* on NiO (20 nm) [46].

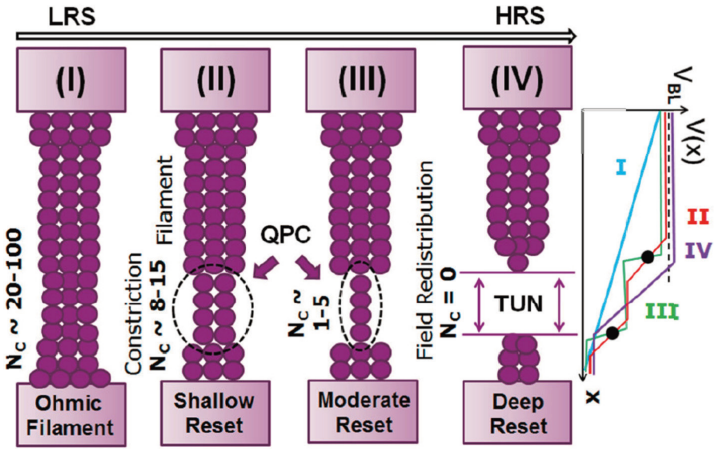


Fig. 11: *The four possible scenarios for the shape and size of the conductive filament (CF) ranging from (I) LRS Ohmic filament to (II) HRS shallow reset, (III) HRS moderate reset with very few defects in the "constriction" to (IV) HRS deep reset, when a tunneling (TUN) barrier is created. The probability to end up in the states (II, III, IV) depends on the dielectric material parameters and vacancy transport properties. The plot at the right end shows the potential drop profile for each of these four scenarios. The HRS states in (II, III) correspond to QPC mode of conduction with majority of the voltage dropping across the two F-C interfaces. Reprinted from [52], Copyright 2013 The Japan Society of Applied Physics.*

Given any measured RTN signal, the first task to be carried out is the identification of the different RTN jumps and finding out which of these jumps correspond to vacancy induced fluctuations and which due to carrier transport based fluctuations. A physical model formulation using the quantum point contact (QPC) model is used in this context. The QPC formulation has been previously proposed by Miranda *et. al.* [48] and Cester *et. al.* [49] to describe the non-linear conduction in the post soft-breakdown regime for high-κ dielectrics with very good fit to the measured I-V data. This formulation has been adapted successfully

to describe the conduction in RRAM for both the HRS and LRS states by Degraeve *et. al.* [32] and Suñé *et. al.* recently [50]. As proposed by Degraeve and co-workers, the filament can be phenomenologically represented as a cluster of vacancies with a narrow "constriction" comprising N_C particles (vacancies). It is this constriction that controls the conductivity of the state. The size of this constriction is governed by N_C and the shape is determined by two parabolic energy band frequency parameters, ω_X and ω_Y where these two quantities qualitatively indicate the length and width of the constriction respectively. The lower the ω_X, the longer the constriction; the lower the ω_Y, the wider the constriction. Using the QPC model, the *I-V* curve for any integer value of N_C can be simulated. To identify vacancy induced RTN effects, the measured *I-V* sweep data can be superimposed on to the *I-V* simulated curves and the data and its multiple jumps can be fit to the QPC simulation model by optimizing the $\{\omega_X, \omega_Y\}$ values. The first voltage level at which a jump in the I-V data occurs from one level of N_C to the adjacent one (either $N_C + 1$ or $N_C - 1$) is classified as the "*disturb*" voltage (V_{DIST}) [51], which is an indicator of the stability of the HRS state against vacancy perturbation induced RTN. Fig.10 illustrates this methodology of V_{DIST} identification. Along with the value of V_{DIST}, the model also provides us with the initial value of N_C. Note here that the underlying assumption is that the filament does not rupture during RESET; instead it only shrinks in size. However, as shown below, this model can be extended to analyze the case of ruptured filaments as well.

The higher the forming / SET compliance (I_{comp}), the shallower the reset is expected to be, as there are more ion-vacancy recombination events needed for a given reset sweep as I_{comp} is increased. From a logical perspective, we would expect the value of V_{DIST} to be higher for lower I_{comp} (deeper reset). However, contrary to our expectation, the deeper the reset (lower I_{READ}), the lower is the measured V_{DIST} value [51]. This is hard to logically interpret, however, it turns out that this is a unique feature of the QPC model as illustrated by Fig.11 [52]. When the filament is not ruptured and the reset gets deeper (implying less number of defects in the constriction), the overall voltage applied across the dielectric redistributes itself to be localized at the filament – constriction (F-C) interfaces only. As a result, with decreasing N_C, the F-C interface width gets sharper and the potential drop (electric field) is more concentrated and locally enhanced. The interesting feature of the QPC is that the field is almost zero anywhere outside the F-C interface and therefore, the immunity to RTN is dependent on the magnitude of the local field at the F-C interface. The deeper the reset (without filament rupture), the higher the local QPC field and therefore, the lower the V_{DIST} value as observed in [51]. From a quantum physics perspective, the potential is concentrated at the F-C interface due to interference of the incident and reflected electron wave functions there [53]. Considering the dependence of V_{DIST} on the ramp rate and using the thermochemical model for defect generation (creating an oxygen vacancy requires bond breakage of Hf-O bonds), it can be estimated that the "time to disturb" (at $V_{READ} = 0.1$V) can be as low as a few milliseconds to as high as a few mega seconds depending on the extent of reset within the QPC regime [52]. Therefore, although deep reset is desired for higher memory window, if the filament does not rupture, then the system will be highly prone to vacancy-induced RTN effects. When operating within the QPC regime, there is always a trade-off involved in the depth of reset and the RTN immunity.

If the filament is able to undergo rupture during the reset process thereby introducing a tunnel barrier in the dielectric for HRS, the disturb trends are completely reversed. As seen in Fig.12 [54], which shows the V_{DIST} - I_{READ} trend for the same device where deep reset was observed for a few cases when I_{comp} is as low as 0.3 μA, the value of V_{DIST} starts to increase again with deeper reset, against the hypothesis of QPC. This is precisely because the QPC ceases to hold true in this regime and we have basically entered the tunnel (TUN) regime with a dielectric barrier across which the potential drops uniformly. Deeper reset corresponds to thicker tunnel

barrier and therefore prolonged disturb time (enhanced RTN immunity). Whether a particular RESET state is in the QPC or TUN regime can be verified by analyzing the $\{\omega_X, \omega_Y, N_c\}$ values. When force-fitting a deep reset I-V curve to the QPC model, we end up with an unrealistically high value of N_c that contradicts the QPC assumption [54]. We therefore have a robust methodology in place to distinguish between QPC and TUN configurations, according to which the immunity to vacancy-induced RTN is determined.

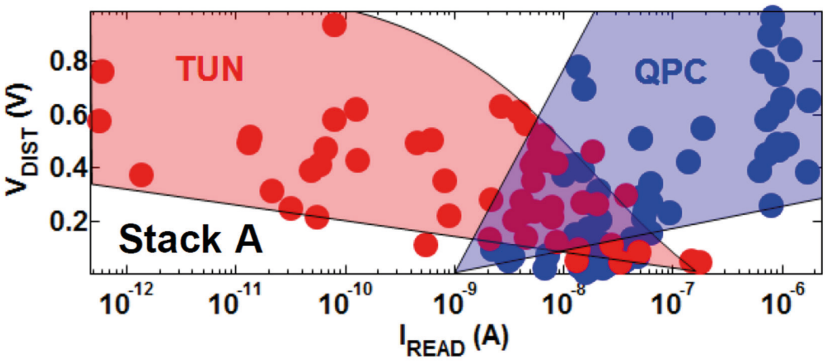


Fig. 12: *Dependence of V_{DIST} on the reset level (I_{READ}) for a very wide range of reset ranging from $I_{READ} = 10^{-12}$ A to 10^{-6} A corresponding to different degrees of soft breakdown (SBD) and progressive breakdown (PBD). The harder the breakdown, the larger is the filament and the lower the chance for rupture (filament remains in QPC mode). © 2013 IEEE. Reprinted, with permission, from [54].*

Considering that RESET is a purely stochastic process that involves recombination of many pairs of oxygen ions and vacancies, we should expect to see a bimodality in the filament configuration which can have a finite non-zero probability of ending up in the QPC or TUN modes. This hypothesis is well confirmed by our analysis of switching for many cycles at different compliance values [54]. Although lower compliance forming and SET enhances the probability of filament rupture (soft breakdown regime), there still exists a finite chance of staying in the QPC regime for the HRS. The vice-versa holds true for the high compliance case (the so-called progressive breakdown regime). Note that the LRS state is not analyzed here because it is more resilient to vacancy perturbations given the uniform potential drop across the whole dielectric for a large size filament (implying low electric field). With the filament configuration being bimodal, we can conclude that the V_{DIST} distribution should also be bimodal.

When the role of the dielectric microstructure is considered, where the dielectric can be amorphous or polycrystalline with grain boundaries (GB), the presence of GB causes a reduction in the V_{DIST} value probably because the GB serves as an easy diffusion path for vacancies to migrate along [55]. Therefore, although the presence of GB may help reduce the forming power, control the variability in the filament size and shape [54] and reduce the bimodality in the filament configuration, it suffers from lower V_{DIST} and has a shallower reset [54] (small memory window) both of which are undesirable.

In summary, we have analyzed the two key mechanisms of RTN in RRAM and identified them to be electron transport based (steady-state) and vacancy perturbation based (non-steady-state). While the former serves as a good defect spectroscopy tool, the latter is the

more critical factor that can disturb the stability of the HRS state. Using the QPC formulation, the vacancy-induced RTN phenomena was studied in-depth for various compliance levels and dielectric material / microstructure. To achieve good switching with robust disturb immunity, it is desirable to have a low LRS state and very deep HRS state (TUN regime). However, it is hard to find dielectric materials that satisfy both these criteria at the same time. From a material design perspective, the field acceleration factor γ (as defined in studies of electrical breakdown [56]]) of the dielectric plays an important role. A high value of γ can simultaneously ensure low forming / SET voltage and a high V_{DIST} value. This factor depends on the relative permittivity and the permanent dipole moment of the high-κ material. Further studies on RTN are essential considering that future devices are being downscaled to areas as low as 10×10 nm^2 [22], where the background noise is relatively low and the RTN effects are expected to be more dominant.

2.5 Read-Disturb and Write-Disturb

A (read) disturb error is an (undesired) change of the programmed state of a memory cell by reading that cell for extended times. More in general, a change of the content of the cell can occur during different memory operation conditions, i.e. not only by multiple reads of the specified bits, but also because of multiple writes or (less likely) a read of other (neighbouring) cells in the same memory array. In the latter case a voltage stress on a non-selected bit may result from directly applied voltages (e.g. on half selected bit or word lines) and/or capacitive coupling phenomena.

The occurrence of neighbouring cell read and write disturbs is strongly dependent on the details of the memory array organization and design. It is clear, however, that the weaker is the isolation between different cells, the more susceptible the array will be to these disturb effects. In particular, they constitute severe design limitations for raw cross-bar RRAM arrays, see [57] for instance.

A general study of disturb susceptibility can be done by applying different voltage stress conditions on a memory cell (with voltage amplitudes that are lower than those required to directly switch the cell).

The intrinsic susceptibility to switching due low voltage stress can be understood from the program voltage versus program pulse width characteristics of an RRAM cell. While measuring this relationship for short pulse widths (spanning the range of normal program pulse widths) indicates a strong non-linearity approaching a kind of threshold voltage for longer pulse times for both SET and RESET switching [58,59], true saturation is not evidenced and switching at lower voltages is still expected, albeit for pulse widths increasing in a super linear way.

If the effect of multiple pulses is the same as of that of one single pulse with a pulse width equal to the sum of the pulse widths, disturb effects can be predicted out of the extrapolation of the measured program voltage versus pulse width behaviour (see Fig.13 [59]) . This assumption is correct only if there are no important transients (e.g of the internal temperature) in the cell during even a single pulse, so that “equilibrium” is reached “instantly”- which may require further extensive studies to possible validate.

For a bipolar switching cell, one would intuitively only expect a possible disturb-caused switching from the cell HRS (RESET) state to the cell LRS (SET) state for voltage stress having the same polarity as the polarity required for SET programming, and, similarly, an LRS to HRS disturb switching only for voltages with the RESET polarity. However, experimentally a decrease of the cell resistance has been observed when applying low voltages pulses with

RESET polarity, see Fig.14 [34]. This effect has been understood by that during the initial (fast) RESET, the cell is not in equilibrium as the cell resistance tends to a "balance value" that is dependent on the applied voltage.

As so, this effect should rather be interpreted as a resistance retention issue with the cell resistance moving to an equilibrium state, and the voltage pulses giving the system the required energy to evolve to that state.

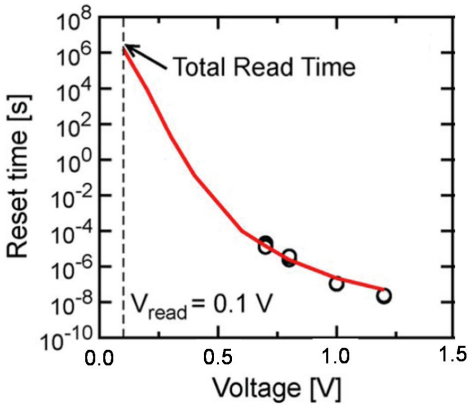


Fig. 13: *Measured and calculated pulse width required for Reset (Reset time) as function of pulse amplitude (Voltage). © 2012 IEEE. Reprinted, with permission, from [59]*

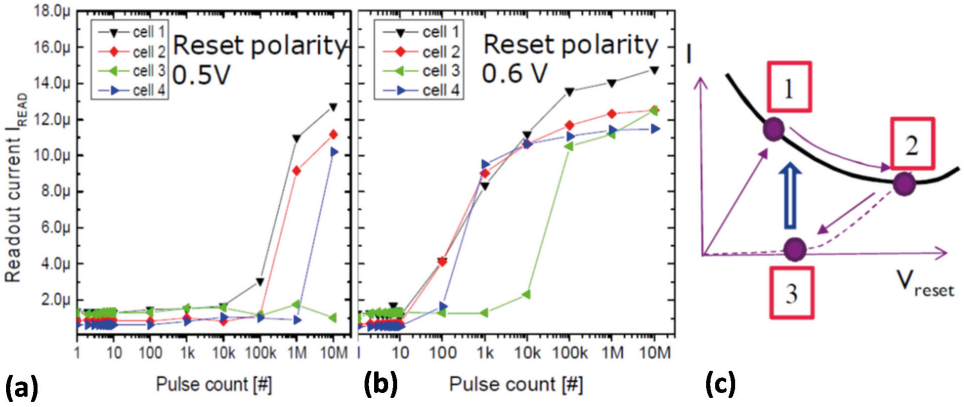


Fig. 14: *Disturb measurements (100ns pulses) at (a) 0.5 and (b) 0.6V in reset polarity. (c) Schematic explanation: (1) transition voltage is reached, 1-->2) reset along dynamicbalance line. 2--> 3) fast down ramp to the OFF state. (3)system tends to return to the ON state (upwards), but because of the low voltage, a time delay is seen (Figs. a en b). © 2012 IEEE. Reprinted, with permission, from [34].*

2.6 Summary

The present section discussed the different reliability aspects of bipolar switching transition metal-oxide RRAM cells. It is shown that insight in the filamentary switching processes on the atomic scale is key to understand the reliability physics. Based on that understanding, material, process and operation conditions may be further tuned to improve the RRAM memory performance towards the targeted application specifications. While different reliability figures (as endurance) have indeed considerably improved, low current operation retention and significant cycle to cycle variability are identified as major concerns for reliable operation of scaled RRAM memories.

Further, we should remain aware that, as the RRAM technology is not yet at the stage of making large density memory arrays in advanced scaled technology, we have today only limited experimental data - mostly on individual cells and/or small arrays. This means that available statistics are still limited, and while we have an understanding on the scale of the "main" bit distributions, we are far from exploring the behaviour of ppm and lower tail bits that may eventually decide on the application of the technology in a real high density memory. This should be further addressed in future reliability studies.

3 Monte-Carlo Circuit Simulation of RRAM Circuits

Both, the evaluation of circuit performance as well as the introduction of effective means in order to actively limit the variability can only be accomplished *in a systematic way* by (i) identifying fundamental (physically based) sources of randomness in the dynamics of RS, and (ii) providing efficient simulation models in order to realize a design space exploration based on fast circuit simulation.

In order to incorporate variability issues into a circuit simulation flow the sources of variability have to be modelled. In this section, modelling is done for a particular class of memristive devices which are based on the Electrochemical Metallization effect (ECM) [61,62]. Since a closed physical device model for ECM cells has been developed recently [63] detailed dynamic simulations including CMOS circuits can be performed. Under the consideration of variability a simple model extension is presented which models a particular aspect of the statistical device behaviour: the random deposition of ions on a filament. The model allows for the simulation of cycle-to-cycle variations and device-to-device variations, and can be applied to larger crossbar arrays using a standard circuit simulator.

3.1 ECM Variability Model for Circuit Simulation

In Fig.15 a sketch of an ECM device based on Cu-SiO_2 is shown. The electrodes are separated by an electronically insulating material (e.g. SiO_2 for Cu-SiO_2 cells) which also acts as an ion conducting layer (electrolyte). If a voltage V_D is applied, electrochemical active ions dissolve into the ion conducting layer and drift towards the counter electrode which is separated from the top electrode by a gap of size S. The deposition of ions on the counter electrode results in the continuous growth of a filament.

The current density for the charge transfer across the electrolyte-electrode interface during the cathodic reduction is described by the Butler-Volmer-equation

$$I_{ion} = A_{fil} \cdot i_0 \cdot \left(e^{\alpha \cdot z\eta_F / V_T} - e^{(\alpha-1)\cdot z\eta_F / V_T} \right) \quad (1)$$

A similar equation holds for the interface A. Since the ionic charge transport is associated with material transport towards the filament, the filament growth is in proportion to the ion current density and is reflected by a shrink of the gap size S

$$\frac{dS}{dt} = -K_p \cdot \frac{I_{ion}}{A_{fil}} \tag{2}$$

In (1) η_F describes the voltage across the interface F, z is the number of charges per ion, V_T the temperature voltage, and α an exchange factor. A_{fil} is the effective active area of the filament, K_p models the relation between electronic charge transport and matter deposition. For small gap sizes S the resistance R_{ion} can be neglected and the voltage across the interfaces (A and F) is approximately half of the device voltage V_D, as the effective areas A_{fil} as well as A_A are almost equal in this case. As soon as S becomes smaller than approximately 0.8 nm the electronic current becomes dominated by a tunneling current which strongly depends on the gap size S [63]. Changes of S by fractions of the atomic diameter can change the conductance G in a significant way. Consequently, the deposition of few atoms on the active filament surface has to be controlled if a particular conductance has to be adjusted up to a given precision.

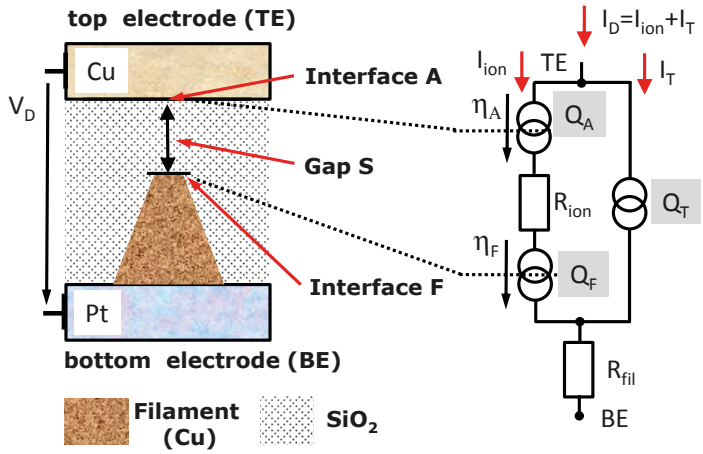


Fig. 15: *Sketch of a device cross section. and Equivalent circuit diagram with nonlinear elements used in SPICE [64].*

The growth of the filament is a discrete process since only integer numbers of atoms are participating in the growth of the filament. The deposition sites as well as the points in time of the deposition are random. An appropriate stochastic model for the discrete ion deposition is given by a Poisson process. The validity of assuming a Poisson process for the description of the filament growth has been shown in [65].

Given an average rate of events - such as the number of depositions on the filament –

$$\lambda = \frac{I_{ion}}{qz} \tag{3}$$

the Poisson distribution predicts the relative frequency of a particular number k of deposition events that can be observed in a given time frame by chance. If exactly k atoms have been deposited the gap S has been effectively decreased by

$$\Delta S = \frac{K_p \cdot qz}{A_{fil}} \cdot k \tag{4}$$

Assuming that the applied device voltage does not change within a short time interval Δt, the mean variation of ΔS becomes a function of time Δt and the filament area:

$$\sigma_{\Delta S} \approx K_p \cdot \sqrt{\frac{\Delta t}{A_{fil}} \cdot q \cdot z \cdot j_0 \cdot \exp\left(\frac{\alpha z \eta_F}{V_T}\right)} \tag{5}$$

From (5) it can be concluded that for scaled devices the variability is seriously affected by the shrink of active filament area A_{fil}.

In order to provide a simulation model which can be executed on a standard circuit simulator (e.g. SPICE [64]), equivalent circuits have to be provided that model the device dynamics by electrical quantities. Here, S is modelled by a voltage V_S (cf. Fig. 16) hold on a capacitance C_S. C_S gets charged or discharged by short current pulses while each pulse represents the deposition (or dissolution) of an individual ion on the filament. Given V_S as well as the device voltage V_D the overall device current I_D=I_{Tunnel}+I_{ion} is obtained by solving the equilibrium state of the voltage-controlled current sources Q_F, Q_A and Q_T. Given the particular I-V characteristics of Q_F, Q_A and Q_T this (standard) operation is automatically done by the circuit simulator. In particular, the voltages η_F as well as η_A are obtained.

The pulse generator (cf. Fig. 16b) generates short pulses which (dis)charge C_S. The interval between adjacent pulses is randomly distributed and the average pulse rate is given by (3) as well as (1). For each pulse the voltage V_S is increased (decreased) by a constant magnitude proportional to (4) using $k=1$. As the deposition rate is a time-dependent function a non-homogeneous Poisson process has to be implemented which tracks the concurrent rate (3). According to the work of Cinlar [66] it is sufficient to consider the *cumulative* event rate function

$$\Lambda(t, t_n) = \int_{t_n}^{t} \lambda(t) dt \tag{6}$$

which counts the average number of events in the intervall $[t_n,t]$. If at time t_n the last deposition event has been observed and a uniformly ditributed random number $RAND$ has been drawn, the next event time t_{n+1} is given by the solution of the equation

$$\Lambda(t_{n+1}, t_n) = -\ln(RAND) \tag{7}$$

which requires to solve (6). Fig. 16c,d show a circuit which realizes the integration, and which represents a detailed implementation of the circuit shown in Fig. 16b. From the drawn random number the signal R_n is derived which is continuously multiplied by the concurrent deposition rate (3). The obtained product gets translated into a current by the source Q_X which charges the capacitance C_X. The voltage V_X across C_X is proportional to (6) and continuously compared to a (constant) reference voltage V_{Tx}. Once V_X exceeds V_{Tx} the capacitance C_X is discharged to 0V, a short pulse V_{pulse} is generated and a new random number is drawn. The quantities V_{Tx}, C_X and the source Q_X are adjusted such that (7) get fulfilled. If V_x reaches V_{Tx}, t_{n+1} has been found with respect to (7) and a pulse on V_{pulse} has been generated at t_{n+1}.

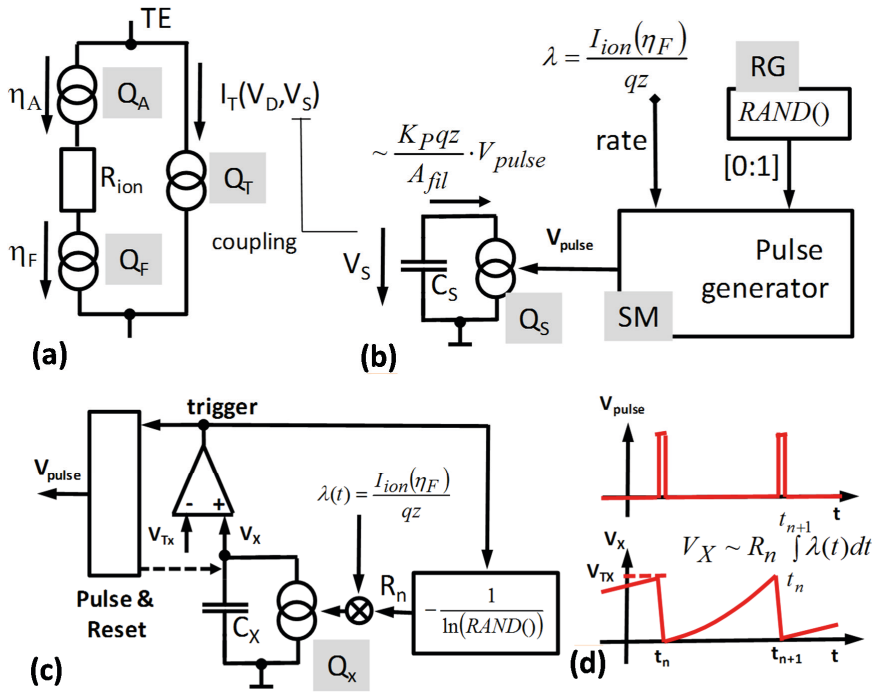


Fig. 16: *(a) ECM device model. The I/V characteristics of Q_A as well as Q_F is given by a Butler Volmer equation similar to (1), Q_T models the tunnel current, which is a function of the device voltage V_D and the gap S; (b) Random pulse generator which models the ion deposition. (c) detailed equivalent-circuit implementation of modules RG and SM, (d) wave forms. The active pulse of the signal V_{pulse} indicates a deposition of an ion on the top of the filament.*

Due to its compatibility with SPICE the presented simulation model can be used in conjunction with CMOS transistor models, e.g. BSIM-CMG (for multi-gate FinFETs). Therefore, it is possible to explore complex circuits comprising a set of active elements (e.g. amplifiers, drivers, coding circuits) as well a passive elements (such as inductive, capacitive, and resistive parasitics, non-linear elements, etc.). In particular, Monte-Carlo simulations have been carried out for various compliance elements [70], crossbars [71], and auxiliary circuits delivering reference voltages for comparator offset compensation [72]. In the following subsection, special focus is set on a Write-Modify algorithm used to program resistive elements in a non-volatile passive crossbar array.

3.2 Simulation of Write-Modify Cycles in Passive Crossbars

The architecture of the passive nanoelectronic crossbar comprising *n* word lines an *m* bi lines is sown in Fig.17. At the crosspoints of word lines and bit lines resistive switches are located. Each resistive switch represents a bit. A logical ‘0’ is represented by a high resistive state (HRS) while a logical ‘1’ is represented by a low resistive state (LRS). If ECM cells are used

as resistive switches the I-V characteristic of each device is almost linear which results is considerable resistive cross talk, also known as sneak paths [67-69]. In [68,69] the optimal setup for the read operation was examined. As a result, if the state of the cell located between word line *j* and bit line *i* is to read, word line *i* has to be tied to ground potential, all other word lines are not connected to a fixed potential, bit line *i* is connected to the read amplifier while other bit lines are tied to the read voltage V_R.

A slightly more stringent condition for the read operation is obtained if word line *j* is tied to ground potential, and all remaining word lines are tied to the read voltage V_R, while bit line *i* is connected to the read amplifier and all remaining bit lines are left unconnected. Here, the read signal is determined by the state of the cells connected to bit line *i* only, see Fig.18b. The output signal in the steady state seen by the read amplifier is then

$$\frac{V_{in}}{V_R} = \frac{\sum_{k=0,k\neq j}^{n+1} G_{i,k}}{G_{i,j} + \sum_{k=0,k\neq j}^{n} G_{i,k}} = \frac{\sum_{k=0,k\neq j}^{n+1} G_{i,k}}{\sum_{k=0}^{n} G_{i,k}} \tag{8}$$

For the following discussion the LRS states as well as the HRS states are randomly distributed among the *n* conductances. When accessing all cells in HSR for all distributions there is a lower bound $V_{in,min}$ which represents a worst case condition for reading a logical one. In turn, when accessing all cells in LSR there is an upper bound $V_{in,max}$, which represents a worst case condition for reading a logical zero. If the LRS is represented by G_{LRS} and the HRS is represented by G_{HRS} these bounds can be specified by considering the worst case condition (9).

$$\frac{V_A}{V_R} = 1 - \frac{1}{n} \leq \frac{Vin}{V_R} < 1 - \frac{1}{G_{LRS}/G_{HRS} + n} \tag{9}$$

The switching threshold V_{TR} of the sense amplifier should be set in between theses bounds in order to maximize the margin between upper bound and lower bound.

However, the (ideal) condition (9) is not guaranteed in practice. If the conductances are subject to variability, upper bound and lower bound approach each other and the margin becomes even tighter.

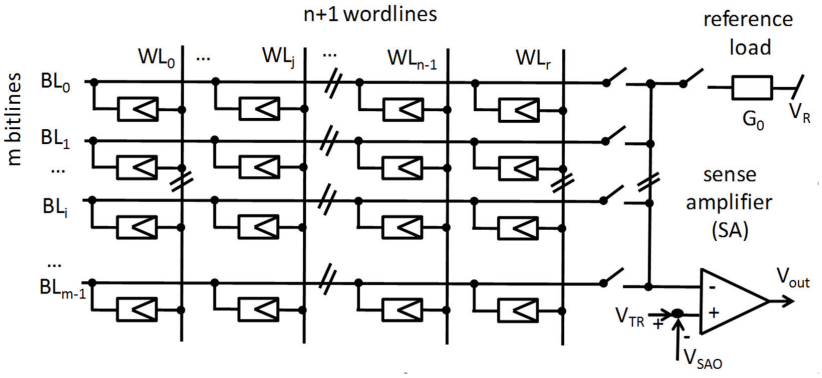


Fig. 17: *Architecture of a passive crossbar comprising n+1 word lines and m bit lines. A cell connected to word line WLX and bit line i is programmed to LRS only if all other cells of bit line i are in HRS.*

For the following discussion it is assumed that G_{LRS} is at least a factor of 10^2 larger than G_{HRS}. For $n < 16$ the upper bound (9) is close to V_R and has less impact on the read performance. Conversely, a worst case situation emerges if all conductances connected to a bit line are in LRS. Then, a read operation results in an array signal V_{in} which is in the range of the lower bound.

The write operation for a single resistive switch is done by applying a sequence of short pulses with fixed voltage level V_P. The accuracy of setting a particular conductance can be adjusted by choosing the pulse width T. The required accuracy is the result of the specification of a defined signal margin for the output signal which has to be kept under all circumstances. For a given signal margin and given parameters the pulse width T is then optimized.

A conductance $G_{i,j}$ gets adapted by applying a short voltage pulse V_P. In order to avoid parasitic programming (i.e. write-disturb) of other cells, bit lines and/or word lines of non-accessed cells have to be set to particular voltages which locally cause smaller device voltages. Fig.18a shows an appropriate configuration, where $V_X=2/3V_P$ and $V_Y=1/3V_P$ holds. Then, the device voltage of non-accessed cells is $1/3V_P$ which significantly reduces the adaptation rate.

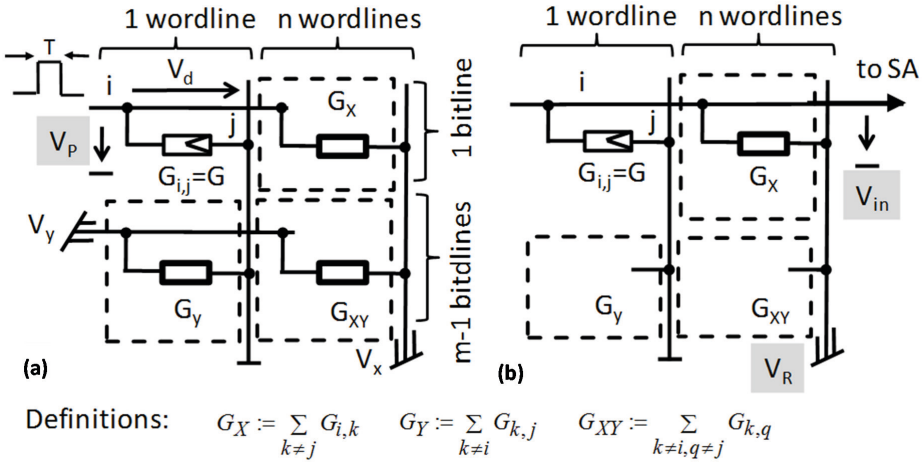


Fig. 18: *a) Equivalent circuit for the write operation; b) Equivalent circuit for the read operation.*

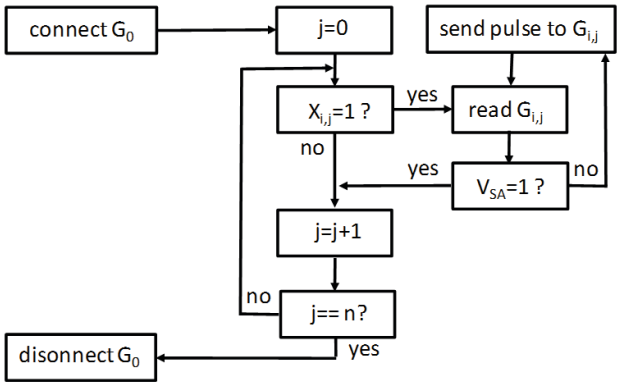


Fig. 19: *Flow of adaptive conductance adaptation for bit line i. $X_{i,j}$ is the state which is to be programmed.*

Since the adaptation effect of a short voltage pulse is only small, many pulses are required to bring a conductance into the right range of magnitude. Under ideal conditions, the conductance should be measured after each adaptation pulse. The adaptation procedure would stop if the appropriate conductance value has been obtained.

Under economical considerations, it is not feasible to measure the exact magnitude of a particular conductance in a passive crossbar. However, some evidence about its relation to other conductances connected to the same bit line can be obtained by assessing the sense amplifier output signal during a subsequent read operation. If the sense amplifier decision is wrong, the accessed conductance is too low and needs some increase which is done by applying a short voltage pulse again. A subsequent read operation assesses whether the increase was sufficiently large and iterates the write/read sequence until the obtained sense amplifier signal is correct.

The main problem involved in this method is, that the conductance is adapted in such a way that the array signal gets close to the sense amplifier threshold if the conductance $G_{i,j}$ is read. Hence, the margin V_M is minimized which is not the desired result. This problem can be solved by introduction of a reference load G_0, which is only active during the read operation in the adaptation process, cf. Fig.17 (upper right corner), Fig.19 shows a complete description of the adaptation process for bit line i.

In the ideal case, the first conductance in the bit line is driven towards

$$G_{i,0} \leftarrow G_0 \cdot \left(\frac{V_R}{V_{TR} - V_{SAO}} - 1 \right) = G_0 \cdot \beta_{TR} \tag{10}$$

where V_R, V_{TR}, V_{SAO} denote the read voltage, ideal sense amplifier threshold voltage, and sense amplifier offset voltage. Once the read operation has succeeded the next resistive switch connected to the bit line is adapted and so forth. For element j the adaptation brings

$$G_{i,j} \leftarrow \left(G_0 + \sum_{k<j} G_{i,k} \right) \cdot \beta_{TR} \tag{11}$$

However, after all n conductances were adapted, they still differ among each other in a significant way as the effective load has increased for each subsequent element j. Hence, the adaptation procedure has to pass through several times. If conductance $G_{i,j}$ is adapted in pass p the condition (12) holds.

$$G_{i,j}(p) \leftarrow \left(G_0 + \sum_{k \neq j} G_{i,k}(p-1) \right) \cdot \beta_{TR} \tag{12}$$

If n weights are to set to LRS along a specified bit line i, the algorithm let the conductances converge towards

$$G_{i,j}(p \to \infty) \leftarrow G_0 \cdot \frac{1}{1/\beta_{TR} - (n-1)} \tag{13}$$

After p iterations the adaptation process stops. Then, for all subsequent read-only operations the conductance G_0 is disconnected from the sense amplifier. Consequently, the array signals (V_{in}) are then smaller (due to the absent load G_0) if a cell in LRS is accessed, and the margin V_M is enlarged.

In fact, the obtained conductances differ from (13) as (i) the increase of the conductance is subject to variability, and (ii) the pulse width T of the programming pulse is finite. Both ef-

fects result in larger conductances than expected, and in variability. The obtained growth of the filament can be divided into two parts by

$$\Delta S = \Delta S_{opt} + \Delta S_{over} \tag{14}$$

where ΔS_{opt} denotes the required (optimal) increase in the filament size and ΔS_{over} the remaining but unnecessary part which results in an over adaptation of G.

If – in the very unlikely event - that in each final adaptation step the condition

$$\Delta S_{opt} << \Delta S_{over} \qquad \Delta S \approx \Delta S_{over} \tag{15}$$

holds, eq. (12) is virtually transformed into

$$G_{i,j}(p) \leftarrow \left(G_0 + \sum_{k \neq j} G_{i,k}(p-1) \right) \cdot e^{\Delta S} \cdot \beta_{TR} \tag{16}$$

which constitutes an effective factor

$$\beta_{TR,eff} := e^{\Delta S} \cdot \beta_{TR} \tag{17}$$

This effective factor substitutes β_{TR} in (12) under worst case conditions.

From (13), the stability of the algorithm is given, if (18) holds:

$$\frac{1}{\beta_{TR,eff}} \overset{!}{>} n-1 \tag{18}$$

The algorithm has – by chance through the distribution of ΔS – the tendency to reduce its stability if the threshold of the sense amplifier is progressively lowered by an offset voltage V_{SAO}, since (18) – by chance - may not hold anymore for large ΔS. It turns out that the described instability results in a larger variability in the conductance distribution, which is notably larger than expected. However, under certain circumstances this instability can be tolerated.

The effects of variability and finite pulse width were examined by Monte Carlo simulations. The number of cells per bit lines were limited to n=2,4,8 and 16 as for larger n the signal margin of the array output signal becomes unrealistically small (note: the read voltage V_R is set to $V_R = 200mV$). Twelve different pulse widths T ranging from 4ns to 1.5µs were used as well as six different filament areas A_{fil} ranging from 1nm^2 to 200nm^2 in order to incorporate the scaling behaviour of the filament growth. The sense amplifier offset was varied between V_A and V_{TR} in twelve steps, and the required margin ratio R_M=V_M/V_R was varied independently between 0.005 and 0.04. The write voltage V_P was set to 0.9V, which is a standard supply voltage for a 28-nm CMOS technology. The conductance G_0 was specified to G_0=1µA/V.

Aim was to find the maximum pulse width T that can be used for robust writing and reliable reading under a particular condition (given n, $A_{Fil,}$) and maintaining a specified margin V_M. As T is maximized (i) the number of subsequent read operations is minimized (which minimizes the overall write time) and (ii) the influence of parasitic effects based on capacitive coupling are minimized. For each parameter conditions 10^4 different simulation runs were performed under random initial conditions for the HRS (approx. 1nA/V).

Fig.20a shows the distribution of conductances after a particular iteration p for n=4,8. After the first iteration the conductance distribution shows an explicit multimodal distribution of conductances while the number of peaks is correlated to n. For larger T, smaller A_{fil} or larger V_{SAO} the distinct peaks tend to overlap and eventually merge, cf. Fig.20b n=4,T=800ns. By all means, after the third iteration (see Fig.20a) the distributions become unimodal, and after the

fourth iteration almost no change in the shape of the distribution was observed, while the centers were only slightly moving towards larger conductances. Therefore, for the subsequent examinations the fourth iteration (p=3) was used as the final iteration loop, and all subsequently obtained results were derived from that particular state.

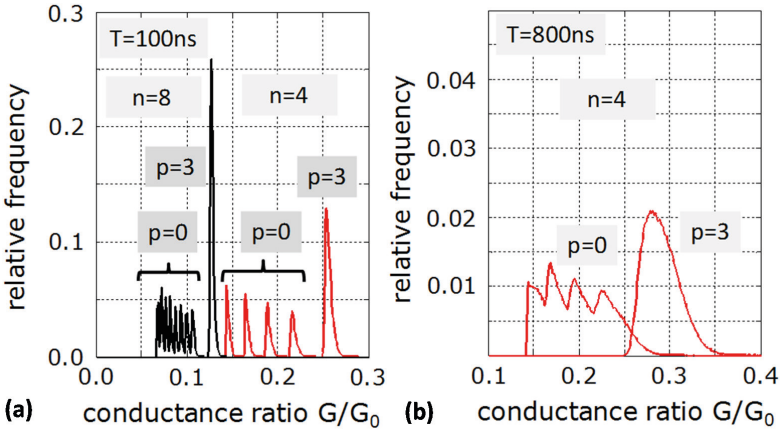


Fig. 20: *Relative frequency of conductance distribution in regard to G_0 for V_R=0.2V, V_P=0.9V, V_{SAO} = 0V, A_{fil}=10nm². a) T=100ns, b) T=800ns.*

The variability of conductances directly influences the output signal of the array in the read mode. After the fourth iteration the reference conductance G_0 was disconnected from the sense amplifier and the worst case condition was determined for the array signal V_{in}. Fig.21a exemplary shows the distribution of the worst case output signals of the array which shows a direct correlation to the conductance distribution.

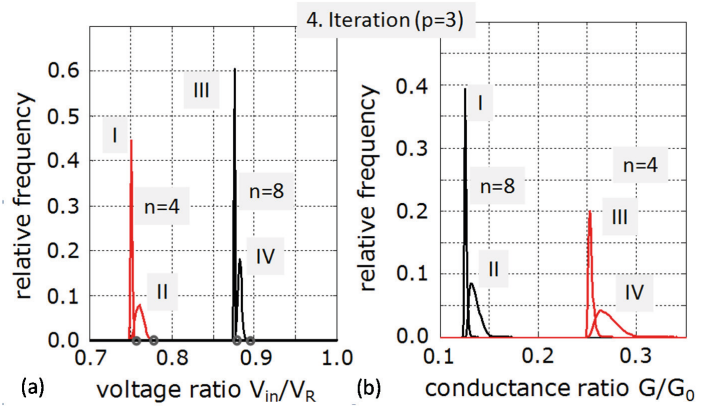


Fig. 21: *(a) Relative distribution of worst-case array voltage V_{in}. Case I,III: T=40ns, case II,IV: T=400ns, A_{fil}=10nm² (b) Relative conductance distribution, p=3,cases I,III: T=40 ns, cases II,IV: T=400ns, A_{fil}=10nm²*

By decreasing the active filament surface the conductance variability increases, cf. eq. (5). For reasonable sense amplifier offset voltages as well as reasonable margins V_M the effect of scaling was determined. As both quantities (V_{SAO}, V_M) were fixed in the subsequent analysis it was possible to determine the prevailing effect that sets the maximum pulse width for writing. The effect of finite pulse width (results in over-adaptation of G) should be reflected in a characteristic that is independent from the filament area A_{fil}. In turn, if a dependency on A_{fil} is found it is evident that the growth statistics has a significant influence on the circuit parameterization.

The dependency of T from the filament surface area A_{fil} is shown in Fig.22 for $n=4,8$, for particular offset voltages, and margins V_M. For large filament surfaces (A_{fil}>40nm^2) T is almost independent from the filament area. Here, the circuit performance depends basically on the finite pulse width T. For smaller filament surfaces the situation changes. Under identical boundary conditions the analysis has shown that the optimization of T results in lower pulse widths in order to maintain the read performance. The discreetness of ion deposition has a significant impact on the variability if a critical filament area is reached.

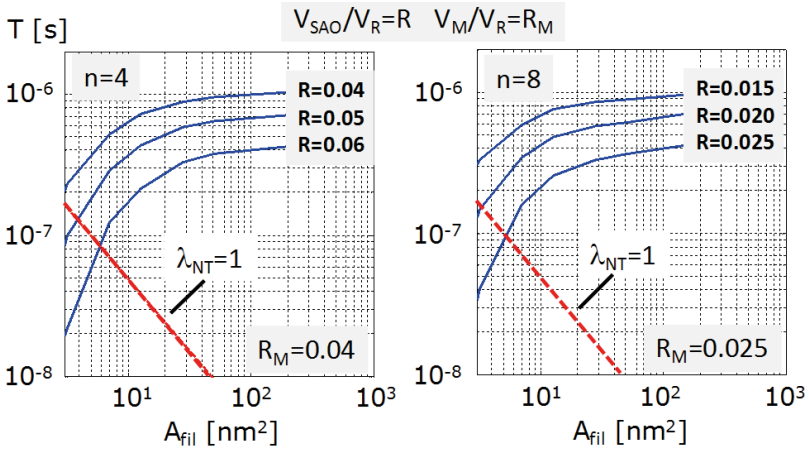


Fig. 22: *Dependency of the pulse width T and the filament area A_{fil}. For reasonable margins V_M and sense amplifier offset V_{SAO} T shows a strong dependence on A_{fil} for A_{fil} < 100nm^2, which indicates that the statistics of the filament growth is prevailing here. For larger filament areas T is saturating which indicates that the finite pulse width is the dominating effect which limits T. The red line denotes the case which provides the deposition of one atom per pulse in the average.*

3.3 Summary

As a summary, the performance of a method used for robustly writing conductive states into resistive switches was analyzed. The resistive switches were located in a passive crossbar array. Here, the focus was set on the cycle-to-cycle variability of the conductance distribution which has a strong impact on the signal margin and hence, the robustness of the circuit. In order to be able to capture the effects of variability an existing device model for ECM cells was extended and prepared to be executable on standard circuit simulator platforms (SPICE).

Under the constraint of a specified signal margin V_M and a specified sense amplifier offset voltage V_{SAO} the pulse width T used for the write pulses was optimized. The results were obtained by extensive Monte Carlo simulations. Although the statistical model is relatively optimistic a significant dependency of the optimization result from the filament area was demonstrated. It definitively shows that any optimization strategy for scaled hybrid circuits has to include the device statistics into the optimization flow and it becomes necessary to simulate devices at the atomic scale.

References

[1] Kim Y-B., Lee S.R., Lee D., Lee C.B., Chang B., Hur J.H., Lee M-J., Park G-S., Kim C.J., Chung U-I., Yoo I-K. and Kim K. (2011) Bi-layered RRAM with Unlimited Endurance and Extremely Uniform Switching. *IEEE 2011 VLSI Technology Symposium (VLSI) Tech. Dig.* , pp.52- 53.

[2] M-J. Lee, C. B. Lee, Lee D., Lee S.R., Chang B., Hur J.H., Kim Y-B., Kim C-J., Seo D.H., Seo S., Chung U-I., I-K. Yoo and Kim K. (2011) A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5−x/TaO2−x bilayer structures. *Nature Materials*, **10**, pp. 625–630,

[3] Lee H.Y., Chen P.S., Wu T.Y., Chen Y.S., Wang C.C., Tzeng P.J., Lin C.H., Chen F., Lien C.H., and Tsai M-J. (2008) Low Power and High Speed Bipolar Switching with A Thin Reactive Ti Buffer Layer in Robust HfO_2 Based RRAM. *IEEE 2008 International Electron Device Symposium (IEDM) Tech Dig.,* pp.297 – 300.

[4] Lee H.Y., Chen Y.S., Chen P.S., Gu P.Y., Hsu Y.Y., Wang S. M., Liu W.H., Tsai C.H., Sheu S.S., Chiang P.C., Lin W.P., Lin C.H., Chen W.S., Chen F.T., Lien C.H., and Tsai M-J. (2010) Evidence and solution of Over-RESET Problem for HfO_X Based Resistive Memory with Sub-ns Switching Speed and High Endurance. *IEEE 2010 International Electron Device Symposium (IEDM) Tech Dig.*, pp.460 – 463.

[5] Chen Y.Y., Govoreanu B., Goux L., Degraeve R., Fantini A., Kar G.S., Wouters D.J., Groeseneken G., J. A. Kittl, Jurczak M. and Altimime L. (2012) Balancing SET/RESET Pulse for $>10^{10}$ Endurance in HfO2 / Hf 1T1R Bipolar RRAM. *IEEE Trans. Electron Dev.*, **59**(12), pp. 3243 – 3249.

[6] Chen Y.Y., Goux L., Clima S., Govoreanu B., Degraeve R., Kar G.S., Fantini A., Groeseneken G., Wouters D.J. and Jurczak M. (2013) Endurance / Retention Trade-off on HfO2/Metal Cap 1T1R Bipolar RRAM. *IEEE Trans. Electron Dev.*, **60**(3), pp.1114 – 1121.

[7] Yang J. J., Zhang M-X., Strachan J.P., Miao F., Pickett M.D., Kelley R.D., Medeiros-Ribeiro G. and Williams R. S. (2010) High switching endurance in TaOx memristive devices. *Appl. Phys. Lett.*, **97**, pp.232102.

[8] Wei Z., Kanzawa Y., Arita K., Katoh Y., Kawai K., Muraoka S., Mitani S., Fujii S., Katayama K., Iijima M., Mikawa T., Ninomiya T., Miyanaga R., Kawashima Y., Tsuji K., Himeno A., Okada T., Azuma R., Shimakawa K., Sugaya H., Takagi T., Yasuhara R., Horiba K., Kumigashira H., and Oshima M. (2008) Highly Reliable TaOx ReRAM and Direct Evidence of Redox Reaction Mechanism. *IEEE 2008 International Electron Device Symposium (IEDM) Tech. Dig*. pp.293-296.

[9] Kawahara A., Azuma R., Ikeda Y., Kawai K., Katoh Y., Tanabe K., Nakamura T., Sumimoto Y., Yamada N., Nakai N., Sakamoto S., Hayakawa Y., Tsuji K., Yoneda S., Himeno A., Origasa K-I., Shimakawa K., Takagi T., Mikawa T., Aono K. (2012) An 8Mb Multi-Layered Cross-Point ReRAM Macro with 443MB/s Write Throughput. *IEEE 2012 international Solid-State Circuits Conference (ISSCC) Tech. Dig.*, pp.432 – 434.

[10] Koveshnikov S., Matthews K., Min K., Gilmer D.C., Sung M.G., Deora S. , Li H.F., Gausepohl S. , Kirsch P.D., Jammy R.(2012) Real-time study of switching kinetics in integrated 1T/ HfOx 1R RRAM: Intrinsic tunability of set/reset voltage and trade-off with switching time. *IEEE 2013 International Electron Device Symposium (IEDM) Tech. Dig.*, pp.486-488.

[11] Chen B., Lu Y., Gao B., Fu Y.H., Zhang F.F., Huang P., Chen Y.S., Liu L.F., Liu X.Y., Kang J.F., Wang Y.Y., Fang Z., Yu H.Y., Li X., Wang X.P., Singh N., Lo G.Q., and Kwong D.L. (2011) Physical Mechanisms of Endurance Degradation in TMO-RRAM. *IEEE 2011 International Electron Device Symposium (IEDM) Tech Dig.*, pp.283 – 286.

[12] Chen Y.Y., Degraeve R., Clima S., Govoreanu B., Goux L., Fantini A., G.S. Kar, Pourtois G., Groeseneken G., D. Wouters, Jurczak M. (2012) Understanding of the Endurance Failure in Scaled HfO2-Based 1T1R RRAM Through Vacancy Mobility Degradation. *IEEE 2012 International Electron Device Symposium (IEDM) Tech. Dig.*, pp. 482 – 485.

[13] Yu S., Guan X., and Wong H-S.P.(2012) Understanding metal oxide RRAM current overshoot and reliability using Kinetic Monte Carlo simulation. *IEEE 2012 International Electron Device Symposium (IEDM) Tech. Dig*., pp. 585 – 588.

[14] Wei Z., Takagi T., Kanzawa Y., Katoh Y., Ninomiya T., Kawai K., Muraoka S., Mitani S., Katayama K., Fujii S., Miyanaga R., Kawashima Y., Mikawa T., Shimakawa K., and K.Aono (2011) Demonstration of High-density ReRAM Ensuring 10-year Retention at 85ºC Based on a Newly Developed Reliability Model. *IEEE 2011 International Electron Device Symposium (IEDM) Tech. Dig.*, pp.721-724.

[15] Wei Z., Takagi T., Kanzawa Y., Katoh Y., Ninomiya T., Kawai K., Muraoka S., Mitani S., Katayama K., Fujii S., Miyanaga R., Kawashima Y., Mikawa T., Shimakawa K., and Aono K. (2012) Retention Model for High-Density ReRAM. *Proceedings IEEE 2012 International Memory Workshop (IMW)*, pp.14 – 17.

[16] Ielmini D., Nardi F., Cagli C., and Lacaita A.L.(2011) Size-dependent retention time in NiO-based resistive switching memories. *IEEE Electron Dev. Lett*. **31**, pp. 353-355.

[17] Yu S., Chen Y.Y., Guan X., Wong H-S.P., and Kittl J.A.(2012) A Monte Carlo study of the low resistance state retention of HfOx based resistive switching memory. *Appl. Phys. Lett*., **100**, pp. 043507.

[18] Wang Y-L., Song Y-L., Yang L-M., Lin Y-Y., Huang R., Zou Q-T., and Wu J-G. (2012) Algorithm-Enhanced Retention Based on Megabit Array of CuxSiyO RRAM. *IEEE Electron Dev. Lett.*, **33**(10), pp.1408 – 1410

[19] Ninomiya T., Wei Z., Muraoka S., Yasuhara R., Katayama K., and Takagi T. (2013) Conductive Filament Scaling of TaOx Bipolar ReRAM for Improving Data Retention Under Low Operation Current. *IEEE Trans. Electron Dev.*, **60**(4), pp. 1384 – 1389.

[20] I.G. Baek, M. S. Lee, Seo S., M. J. Lee, D. H. Seo, D.-S suh, J. C. Park, S. O. Park, H. S. Kim, I. K. Yoo, U.-I Chung, and J. T. Moon (2004) Highly scalable nonvolatile memory using simple binary oxide driven by asymmetric pulses. *IEEE 2004 International Electron Device Symposium (IEDM) Tech. Dig*., pp. 587-590.

[21] Govoreanu B., Kar G.S., Y. –Y Chen, V. Parashiv, S. Kubicek, Fantini A., I. P. Radu, Goux L., Clima S., Degraeve R., N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, Pourtois G., H. Bender, Altimime L., Wouters D.J., Kittl J.A.and M. Jurckzack (2011) 10x10nm^2 Hf/HfO Crossbar resistive RAM with excellent performance, reliability and low-energy operation. *IEEE 2011 International Electron Device Symposium (IEDM) Tech Dig*., pp 792-732.

[22] Vandelli L., Padovani A., Larcher L., Bersuker G., D. Gilmer and P. Pavan (2011) Modeling of the forming operation in HfO2-based resistive switching memories. *Proceedings IEEE 2011 International Memory Workshop (IMW)*, p.1-4.

[23] N. Raghavan , Fantini A., Degraeve R., P.J. Roussel, Goux L., Govoreanu B., Wouters D.J., Groeseneken G. and Jurczak M. (2013) Statistical insight into controlled forming and forming free stacks for HfOx RRAM. *Microelectronic Engineering*, **109**, pp 177–181.

[24] Degraeve R., Goux L., Clima S., Govoreanu B., Chen Y.Y., G.S. Kar, P.J. Roussel, Pourtois G., Wouters D.J., Altimime L., Jurczak M., Groeseneken G. and J.A. Kittl (2012) Modeling and tuning the filament properties in RRAM metal oxide stacks for optimized stable cycling. *Proceedings 2012 International Symposium on VLSI Technology, Systems, and Applications (VLSI-TSA),* pp.1-2.

[25] BJ Choi, AC Torrezan, KJ Norris, F Miao, JP Strachan, MX Zhang (2013), Electrical performance and scalability of Pt dispersed SiO2 nanometallic resistance switch, Nano letters 13 (7), 3213-3217.

[26] J. Liang, S. Yeh, S. Wong, and H. -S. Philip Wong (2012) Scaling Challenges for the Cross-point Resistive Memory Array to Sub-10nm Node – An Interconnect Perspective. *Proceedings IEEE 2012 International Memory Workshop (IMW)*, pp 61-64.

[27] L. Zhang, S. Cosemans, D. J.Wouters, Govoreanu B., Groeseneken G., Jurczak M. (2013) Analysis of vertical Cross-Point Resistive Memory (VRRAM) for 3D RRAM Design. *Proceedings IEEE 2013 International Memory Workshop (IMW)*,, pp 155-158

[28] S. C. Chae, J. S. Lee, S. Kim, S. B. Lee, S. H. Chang, C. Liu, B. Kahng, H. Shin, D.-W. Kim, C. U. Jung, Seo S., M. –J. Lee and T.W. Noh (2008) Random circuit breaker network model for unipolar resistance switching. *Adv. Mater.,* **20**(6), pp. 1154-1159.

[29] S. B. Lee, J. S. Lee, S. H Chang, H. K. Yoo, B. S. Kang, B. Kahng, Lee M-J., Kim C.J. and T. W. Noh, (2011) Interface-modified random circuit breaker network model applicable to both bipolar and unipolar resistance switching. *Appl. Phys. Lett.*, **98**, pp. 033502.

[30] Yu S., Guan X. and Wong H-S.P.(2012) On the switching parameter variation of metal oxide RRAM-Part II: Model corroboration and device design strategy. *IEEE Trans. Electron Dev.*, **59**(4), pp. 1183-1188.

[31] Bersuker G., Gilmer D.C., Veksler D., Kirsch P., Vandelli L., Padovani A., Larcher L., McKenna K., Shluger A., V. Iglesias, Porti M., and Nafria M. (2011) Metal oxide resitive memory switching mechanism based on conductive filament properties. *J. Appl. Phys*., **110**(12), 2011

[32] Degraeve R., Roussel P., Goux L., Wouters D.J., J Kittl, L. Altimime, Jurczak M. and u(2012) Generic learning of TDDB applied to RRAM for improved understanding of conduction and switching mechanism through multiple filaments. *IEEE 2010 International Electron Device Symposium (IEDM) Tech. Dig*., pp 632-635.

[33] Ielmini D. (2011) Modeling the universal set/reset characteristics of bipolar RRAM by field- and temperature-driven filament growth. *IEEE Trans. Electron Dev.,* **58**(12), pp. 4309-4317.

[34] Degraeve R., Fantini A., Clima S., Govoreanu B., Goux L., Chen Y.Y., Wouters D.J., Roussel P., Kar G.S., Pourtois G., Cosemans S., Kittl J.A., Groeseneken G., Jurczak M., and Altimime L. (2012) Dynamic 'hour glass' model for SET and RESET in HfO2 RRAM, *IEEE 2012 VLSI Technology Symposium (VLSI) Tech. Dig*., pp. 75-76.

[35] Fantini A., Goux L., Degraeve R., Wouters D.J., Raghavan N., G. Kar, Belmonte A., Chen Y.Y., Govoreanu B. and Jurczak M. (2012) Intrinsic switching variability in HfO2 RRAM. *Proceedings IEEE 2012 International Memory Workshop (IMW)*,, pp. 45-48.

[36] Balatti S., Ambrogio S., Gilmer D.C., Ielmini D. (2013) Set variability and failure Induced by complementary switching in bipolar RRAM. *IEEE Electron Dev. Lett*., **34**(7), pp. 861-863.

[37] Chen Y.S., Lee H.Y., Chen P.S., Gu P.Y., Chen C.W., Lin W.P., Liu W.H., Hsu Y.Y., Sheu S.S., Chiang P.C., Chen W.S., Chen F.T., Lien C.H. and Tsai M.J. (2009) Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity. *IEEE 2009 International Electron Device Symposium (IEDM) Tech.Dig.,* pp. 105-108.

[38] Terai M., Sakotsubo Y., Saito Y., Kotsuji S. and Hada H. (2010) Memory-state dependence of random telegraph noise of Ta_2O_5/TiO_2 stack ReRAM. *IEEE Electron Dev. Lett.,* **31**(11), pp. 1302-1304.

[39] Lee D., Lee J.; Jo M., Park J., Siddik M. and Hwang H. (2011) Noise-Analysis-Based Model of Filamentary Switching ReRAM With ZrO_x / HfO_x Stacks. *IEEE Electron Dev. Lett.,* **32**(7), pp. 964-966.

[40] Lee J.K., Lee J.W., Park J., Chung S.W., Roh J.S., Hong S.J., Cho I.W., Kwan H.I. and Lee J.H. (2011) Extraction of trap location and energy from random telegraph noise in amorphous TiO_x resistance random access memories. *Appl. Phys. Lett.*, **98**, pp.143502.

[41] Tseng Y.H., Shen W.C., Huang C.E., Lin C.J. and King Y.C. (2010) Electron trapping effect on the switching behavior of contact RRAM devices through random telegraph noise analysis. *IEEE 2010 International Electron Device Symposium (IEDM) Tech.Dig.,* pp.28.5.1 - 28.5.4.

[42] Raghavan N., Degraeve R., Fantini A., Goux L., S. Strangio, Govoreanu B., Wouters D.J., Groeseneken G. and Jurczak M. (2013) Microscopic origin of random telegraph noise fluctuations in aggressively scaled RRAM and its impact on read disturb variability. *IEEE 2013 International Reliability Physics Symposium (IRPS)*, Monterey, California, pp.5E.3.1-5E.3.7.

[43] Soni R., Meuffels P., Petraru A., Weides M., Kügeler C., Waser R. and Kohlstedt H. (2010) Probing Cu doped $Ge_{0.3}Se_{0.7}$ based resistance switching memory devices with random telegraph noise. *J. Appl. Phys.*, **107**, 024517.

[44] Puglisi F.M., Pavan P., Padovani A., Larcher L. and Bersuker G. (2010) Random telegraph signal noise properties of HfO_x RRAM in high resistive state. *Proceedings of the IEEE 2012 European Solid-State Device Research Conference (ESSDERC),* pp.274-277.

[45] Veksler D., Bersuker G., Chakrabarti B., Vogel E., Deora S. ,Matthews K., Gilmer D.C., Li H.F., Gausepohl S. and Kirsch P.D. (2012) Methodology for the statistical evaluation of the effect of random telegraph noise (RTN) on RRAM characteristics. *IEEE 2012 International Electron Device Symposium (IEDM) Tech.Dig.*, pp.219 – 222.

[46] Ielmini D., Nardi F. and Cagli C.(2010) Resistance-dependent amplitude of random telegraph-signal noise in resistive switching memories/ *Appl. Phys. Lett.*, **96**, pp. 053503.

[47] Lee J.K., Jeong H.Y., Cho I.T., Lee J.Y., Choi S.Y., Kwon H.I. and Lee J.H. (2010) Conduction and Low-Frequency Noise Analysis in Al / TiO_x / Al Bipolar Switching Resistance Random Access Memory Devices. *IEEE Electron Dev. Lett.,* **31**(6), pp. 603-605.

[48] Miranda E. and Suñé J. (2001) Analytic modeling of leakage current through multiple breakdown paths in SiO_2 films, *Proceedings IEEE 2001 International Reliability Physics Symposium (IRPS),* pp. 367-379.

[49] Cester A., Bandiera L., Suñé J., Boschiero L., Ghidini G. and Paccagnella A. (2001) A novel approach to quantum point contact for post soft breakdown conduction. *IEEE 2001 International Electron Device Symposium (IEDM) Tech. Dig.,* pp.305-308.

[50] Lian X., Long S., Cagli C., Buckley J., Miranda E., Liu M. and Suñé J. (2012) Quantum point contact model of filamentary conduction in resistive switching memories, *Proceedings 13th 2012 International Conference on Ultimate Integration on Silicon (ULIS),* pp. 101-104.

[51] Raghavan N., Degraeve R., Fantini A., Goux L., Wouters D.J., Groeseneken G. and Jurczak M. (2013) Modeling the impact of reset depth on vacancy induced filament perturbations in HfO_2 RRAM. *IEEE Electron Dev. Lett.*, **34**(5), pp.614-616.

[52] Raghavan N., Degraeve R., Goux L., Fantini A., Wouters D.J., Groeseneken G. and Jurczak M. (2013) RTN insight to filamentary instability and disturb immunity in ultra-low power switching HfO_x and AlO_x RRAM. *IEEE 2013 VLSI Technology Symposium (VLSI) Tech.Dig.,* T163-T164.

[53] Ulreich S. and Zwerger W. (1998) Where is the potential drop in a quantum point contact?, *Superlattices and Microstructures*, **23**(3–4), pp. 719-730.

[54] Raghavan N., Degraeve R., Fantini A., Goux L., Wouters D.J., Groeseneken G. and Jurczak M. (2013) Stochastic variability of vacancy filament configuration in ultra-thin dielectric RRAM and its impact on OFF-state reliability. *IEEE 2013 International Electron Device Symposium (IEDM) Tech.Dig.* pp. 554-557.

[55] McKenna K. and Shluger A. (2009) The interaction of oxygen vacancies with grain boundaries in monoclinic HfO_2. *Appl. Phys. Lett.,* **95**, pp. 222111.

[56] McPherson J.W. and Mogul H. C. (1998) Underlying physics of the thermochemical E model in describing low-field time-dependent dielectric breakdown inSiO2 thin films. *J. Appl. Phys.* **84**(3), pp.1513-1523.

[57] Chen A. (2011) Accessibility of nano-crossbar arrays of resistive switching devices. *Nanotechnology (IEEE-NANO),* Aug.2011, pp.1767-1771.

[58] Schindler C., Staikov G., and Waser R., (2009) Electrode kinetics of Cu-SiO2-based resistive switching cells: Overcoming the voltage-time dilemma of electrochemical metallization memories. *Appl. Phys. Lett.* **94**, pp. 072109.

[59] Larentis S., Nardi F., Balatti S., Gilmer D. C., and Ielmini D. (2012) Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part II: Modeling. *IEEE Trans. Electron Dev.*,**59**(9), p. 2468-2475.

[60] C.Zambelli, A.Chimenton, P.Olivo, "Reliability Issues of NAND Flash Memories", in "Inside NAND Flash Memories", R.Micheloni, L.Crippa, A.Marelli (eds.), Springer, 2010

[61] R.Waser, "Electrochemical and Thermochemical Memories", IEDM, pp 1-4, 2008.

[62] Waser, R., Dittmann, R., Staikov, G. and Szot, K. (2009), Redox-Based Resistive Switching Memories – Nanoionic Mechanisms, Prospects, and Challenges. Adv. Mater., 21: 2632–2663.

[63] S.Menzel, U.Böttger, and R.Waser, "Simulation of Multilevel Switching in Electrochemical Metallization Memory Cells", Journal of Applied Physics Vol. 111, pp.014501-014501-5, Jan. 2012

[64] Nagel, L. W, and Pederson, D. O., *SPICE (Simulation Program with Integrated Circuit Emphasis)*, Memorandum No. ERL-M382, University of California, Berkeley, Apr. 1973

[65] R.Soni, P.Meuffels, G.Staikov, R.Weng, C.Kügeler, A.Petraru, M.Hambe, R.Waser, H.Kohlstedt, "On the stochastic nature of resistive switching in Cu doped Ge0.3Se0.7 based memory devices", Journal of Applied Physics 110, 054509, 2011

[66] E.Cinlar, "Introduction to Stochastic Processes". Prentice Hall, 1975

[67] E.Linn, R.Rosezin, C.Kügeler, R.Waser, "Complementary Resistive Switches for Passive Nanocrossbar Memories", Nature Materials 9, pp 403-406, 2010

[68] A.Flocke, T.G.Noll, C.Kügeler, C.Nauenheim, R.Waser. "A Fundamental Analysis of Nano-Crossbars with Non-Linear Switching Materials and its Impact on TiO2 as a Resistive Layer". 8th IEEE Conference on Nanotechnology, pages 319–322, 2008.

[69] A.Flocke, T.G. Noll, "Fundamental Analysis of Write-Operations in Resistive Crossbar Arrays", nanoelectronic days 2008

[70] A.Heittmann, T.G. Noll, Variability Evaluation of Feedback Circuits Used in Nanoelectronic Memristive/CMOS Circuits", Great Lakes Symposium VLSI (GLSVLSI), Paris, France, pp. 137-142, 2013

[71] A.Heittmann, T.G.Noll, "A Monte Carlo Analysis of a Write Method used in Passive Nanoelectronic Crossbars", Procs. of ACM/IEEE Nanoarch, pp.93-100, **2012**

[72] A.Heittmann, T.G. Noll, "Variability Analysis of a Hybrid CMOS/RS Nanoelectronic Calibration Circuit", ISCAS 2014, pp.1656-1659, 2014

E 2 Ultimate Physical Limit of Scaling

Victor V. Zhirnov
Semiconductor Research Corporation, USA

Contents

Contents **1**

1 Introduction **2**

2 Basic operations of ICT devices **2**

3 Essential Physics of ICT devices **4**
3.1 The Use of Particles to Represent Binary Information 4
3.2 Example I: Minimum Energy of Computing 7
3.3 Example II: Scaling Limits of Charge-Based Memory 7
3.4 Spintronic ICT 10
3.5 Scaling below 5 nm 12

4 Scaling limits of nanoionic devices **13**
4.1 Switching mechanisms and the material systems 13
4.2 Atomic filament: Classical and Quantum resistance 14
4.3 Conductance in the presence of barriers 17
4.4 Barriers in atomic gaps 18
4.5 Transmission through atomic gaps 21
4.6 Interface Controlled Resistance (ICR) 22
4.7 Stability of the minimal nanoionic state 28
4.8 Switching speed and energy of ultimate nanoionics devices 31

5 Summary **32**

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

Device scaling and energy consumption during computation has become a matter of strategic importance for modern Information and Communication Technologies (ICT). The central question addressed in this chapter is: What is the smallest volume of matter needed for ICT devices, such as memory or logic?

This chapter considers physical principles and trends in nano-scale information processing devices. It provides a coherent description and scaling analysis of diverse nanoelectronic devices that operate in different physical domains. A generic device abstraction for computational elements is developed for a uniform treatment of several classes of nanodevices that use different information carriers, such as, electrons, spins or atoms/ions. Estimates of theoretically attainable geometric scaling limits for these diverse devices are also given. Theoretical feasibility of the 1-nm devices will be justified based on electrical properties of the few-atom systems and an expository, physics-based, framework for the estimation of physical electrical and thermal limits for atomic contacts and interfaces, with an emphasis on nanoionics memories will be provided.

2 Basic operations of ICT devices

Information can be defined as a quantitative measure of distinguishability of a physical subsystem from its environment [1]. Information of arbitrary kind and amount can be represented by combination of just two distinguishable states (known as binary states and often marked as state 0 and state 1). A system with two distinguishable states forms the basic ICT element (binary switch or memory cell) shown in Fig. 1. It consists of: 1) two states 0 and 1, which are equally attainable and distinguishable; 2) a means to control the change of the state (WRITE operation); 3) a means to detect the state (READ operation); and/or 4) a means to communicate with other binary switches (TALK operation).

Information-processing systems represent system states in terms of physical variables. One way to create physically distinguishable states is by the presence or absence of material particles in a given location. Examples are electrons in transistor gates, or atoms/ions in nanoionic devices. Essential requirements for the implementation of an ICT device are the ability to move the particles from state 0 to state 1 and from state 1 to state 0, when an external WRITE signal is applied, and conversely, the particles must remain in its position for a sufficiently long time. For example a typical practical requirement for a nonvolatile memory element is the state lifetime (the retention time) $t_r \sim 10$ years$\sim 3\times10^8$ s.

In low-energy ICT systems the operating voltage has to be decreased, and the most advanced logic circuits currently operate at V=0.8-1 Volt. The operating frequency of the current baseline logic circuits is in the GHz range which equates to switching speed of <1ns. It is desirable that the operating voltage and the speed of operation of memory be in the same range as logic (for the reasons of energy saving and operational compatibility), and therefore one needs to detect the presence/absence of the particles in e.g., the state 1 by a fast (~ns) READ operation. In a typical READ operation, a read voltage V_r is applied to a communication line and a voltage V_{sense} is sensed at the load resistor R_L (see Fig. 2) and compared to a reference voltage. The discrete nature of electrical charge sets a fundamental restriction on the minimal current needed for a fast and reliable reading. The read current passing through the load resistor in Fig. 2 is:

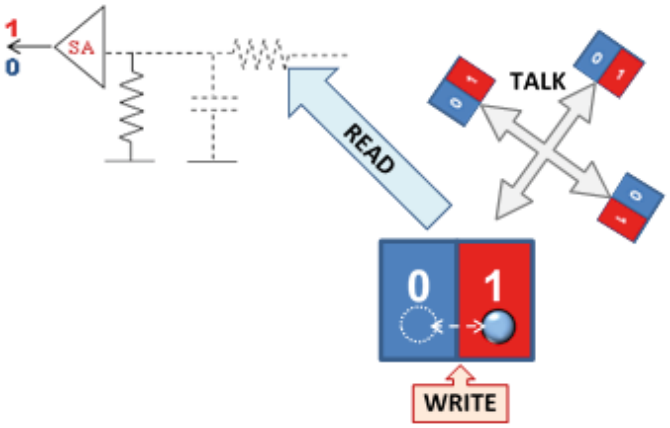


Fig. 1: *An abstract binary ICT element.*

$$I_r = \frac{q}{t_r} = \frac{Ne}{t_r} \tag{1a}$$

and the minimal current, when only one electron passes through the resistor during the read interval is

$$I_{r_{\min}} = \frac{e}{t_r} \tag{1b}$$

For example if the specified read time is t_r ~1ns, the minimal read current is ~100pA, which corresponds to only one electron (on average) passing through the load resistor R_L during the read interval. For such small numbers, the inevitable statistical fluctuations can easily lead to an erroneous result. The limits of READ operation therefore can be assessed based on the 'margin of error' for each state: if N is the average number of electrons, injected in the READ circuit during the read interval t_r, the margin of error is given by the Poisson distribution as a standard deviation of $\approx \pm\sqrt{N}$, and the corresponding relative error can be estimated as:

$$\delta \sim \frac{\sqrt{N}}{N} = \frac{1}{\sqrt{N}} \tag{2}$$

For example, if N=1, the error δ=100%, and for a reliable reading larger number of electrons is needed. From (1) and (2):

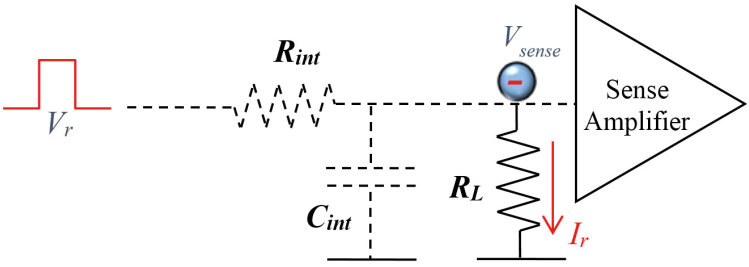


Fig. 2: *A typical READ operation in an ICT element.*

$$\delta \sim \sqrt{\frac{e}{I_r t_r}} \tag{3}$$

For example, if the specified read time and error margin are respectively t_r~1ns and δ ~1%, the read current should be >1μA. In addition, to distinguish between two informational states (e.g. ON and OFF), the ratio of the currents for the two states should be at least ten [2].

In summary, the essential properties of an advanced high-performance ICT device are:

Operating Voltage: ~ 1 V

Read Time: ~ 1ns

Read Current: >1μA

ON/OFF Ratio >10

State lifetime: ~3×10^8 s (nonvolatile memory applications)

3 Essential Physics of ICT devices

Three essential properties of a binary switch are Distinguishability, Controllability and Communicativity. We say that a binary switch is Distinguishable if and only if the binary state (0 or 1) can be determined with an acceptable degree of certainty by a measurement (READ operation). The binary switch is Controllable if an external stimulus can reliably change the state of the system from 0 to 1 or from 1 to 0. (WRITE operation) The binary switch is communicative if it is capable of transferring its state to other binary switches (TALK operation).

3.1 The Use of Particles to Represent Binary Information

Information-processing systems represent system states in terms of physical variables. One way to create physically distinguishable states is by the presence or absence of material particles or fields in a given location. Fig. 3a shows an abstract model for a binary switch whose state is represented by different positions of a material particle. In principle, the particle can possess arbitrary mass, charge etc. The only two requirements for the implementation of a particle-based binary switch are the ability to detect the presence/absence of the particle in e.g., the location x_1, and the ability to move the particle from x_0 to x_1 and from x_1 to x_0 . Let $\Pi_{correct}$ be the probability that the binary switch is in the correct state at an arbitrary time after the command to achieve that state is given. Alternatively, one can use the probability of error $\Pi_{err} = 1 - \Pi_{correct}$. A necessary condition for the distinguishability of a binary switch is

$$\Pi_{correct} > \Pi_{err} \tag{4a}$$

Or equivalently:

$$\Pi_{err} < 0.5 \tag{4b}$$

As it will be discussed below, in physical realizations of binary switches, there always is some error probability ($\Pi_{err}>0$) in the operation of the switch. Since the error probability

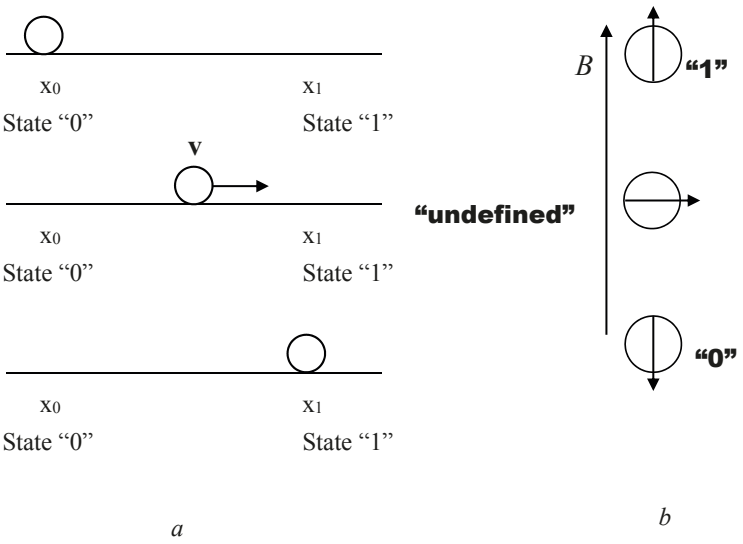


Fig. 3: *An abstract model for operation of a binary switch formed (a) by different locations of material particles and (b) by opposite direction of electron spin magnetic moment in external magnetic field*

cannot exceed 0.5, in the following analysis we will use the condition (4b) to estimate parameters of a binary switch in the limiting case.

Consider again, a binary switch where the binary state is represented by particle location (Fig. 4a). Until now, it was assumed that the information-defining particle in the binary switch has zero velocity/kinetic energy, prior to a WRITE command. However, each material particle at equilibrium with the environment possesses kinetic energy of ½ $k_B T$ per degree of freedom due to thermal interactions, where k_B is the Boltzmann's constant and T is temperature. The permanent supply of thermal energy to the system occurs via mechanical vibrations of atoms (phonons) and via the thermal electromagnetic field of photons (background radiation). In order to prevent the location of the particle from changing randomly due to thermal excitation, energy barriers needs to be constructed that limit particle movements. The energy barrier, separating the two states in a binary switch is characterized by its height E_b and width a (Fig. 4b). The barrier height, E_b, must be large enough to prevent spontaneous transitions (errors). Two types of unintended transitions can occur: "classical" and "quantum". The "classical" error occurs when the particle jumps over barrier. This can happen if the kinetic energy of the particle E is larger than E_b. The corresponding probability for over-barrier transition Π_C (referred herein as "classic" error probability), is obtained from the Boltzmann distribution as:

$$\Pi_C = \exp\left(-\frac{E_b}{k_B T}\right) \tag{5}$$

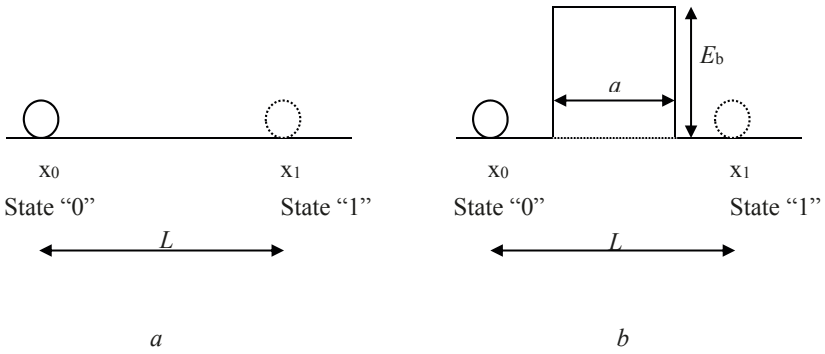


Fig. 4: *Illustration of an energy barrier to preserve the binary states*

Another class of errors, called "quantum errors", occur due to quantum mechanical tunneling through the barrier of finite width a. If the barrier is too narrow, spontaneous tunneling through the barrier will destroy the binary information. The conditions for significant tunneling can be estimated using the Heisenberg uncertainty principle; as is often done in the texts on the theory of tunneling [3]:

$$\Delta x \Delta p \geq \frac{\hbar}{2} \tag{6}$$

The uncertainty relation (6) can be used to estimate the limits of distinguishability. Consider again a binary device with a barrier in Fig. 4b. As is known from Quantum Mechanics, a particle can pass (tunnel) through a barrier of finite width even if the particle energy is less than the barrier height, E_b. An estimate of how thin the barrier must be to observe tunneling can be made from (6). For a particle at the bottom of the well, the uncertainty in momentum is $\sqrt{2mE_b}$, which gives:

$$\sqrt{2mE_b}\Delta x \approx \frac{\hbar}{2} \tag{7}$$

Eqn. (7) states that by initially setting the particle on one side of the barrier, one can find the particle on either side with high probability, if Δx is of the order of the barrier width a. That is, the condition for losing distinguishability is $\Delta x \geq a$, and the minimum barrier width is:

$$a_{\min} = a_H \approx \frac{\hbar}{2\sqrt{2mE_b}} \tag{8}$$

a_H is the Heisenberg distinguishability length for "classic to quantum transition".

For $a < a_H$, tunneling probability is significant, and therefore particle localization is not possible. To estimate the probability of tunneling, we re-write (8), taking into account the tunneling condition $a \leq \Delta x$:

$$\sqrt{2m}(a\sqrt{E_b}) \leq \frac{\hbar}{2} \tag{9a}$$

From (9a), we can also write the "tunneling condition" in the form

$$1-\frac{2\sqrt{2m}}{\hbar}a\sqrt{E_b}\geq 0 \tag{9b}$$

Since for small x, $e^{-x} \sim 1-x$, the tunneling condition then becomes

$$\exp\left(-\frac{2\sqrt{2m}}{\hbar}\cdot a\cdot\sqrt{E_b}\right)\geq 0 \tag{9c}$$

The left side of Eqn. (9c) has the properties of probability. Indeed, it represents the tunneling probability through a rectangular barrier given by the Wentzel-Kramers-Brillouin(WKB) approximation [4]:

$$\Pi_{WKB}\sim\exp\left(-\frac{2\sqrt{2m}}{\hbar}\cdot a\cdot\sqrt{E_b}\right) \tag{10}$$

This equation also emphasizes the parameters controlling the tunneling process. They are the barrier height E_b and barrier width a as well as the mass m of the information-bearing particle. If separation between two wells is less than a, the structure of Fig. 4b would allow significant tunneling.

3.2 Example I: Minimum Energy of Computing

The minimum energy of binary transition is determined by the energy barrier. The work required to suppress the barrier is equal or larger than E_b. Thus, the minimum energy of binary transition is given by the minimum barrier height in binary switch. The minimum barrier height can be found from the distinguishability condition (5), which requires that the probability of errors $\Pi_{err}<0.5$. In this simple example we consider the case when only "classic" (i.e. thermal) errors can occur, i.e. Π_{err} is given by (5). Solving (5) for $\Pi_{err}=0.5$, obtain the Boltzmann's limit for the minimum barrier height, E_{bB}:

$$E_{bB}=k_BT\,ln\,2\approx 0.69k_BT=2.87\cdot 10^{-21}J \quad (T=300\text{K}) \tag{11}$$

Eq. (11) corresponds to the minimum barrier height, the point at which distinguishability of states is completely lost due to thermal over-barrier transitions. These transitions represent the thermal (Nyquist-Johnson) noise. In deriving (11), tunneling was ignored, i.e. the barrier width is assumed to be very large, $a>>a_H$ (8).

3.3 Example II: Scaling Limits of Charge-Based Memory

The current baseline memory technologies (DRAM, SRAM, and flash) are based on storing electron charge in a storage node. Two distinguishable states 0 and 1 are created by the presence (e.g. state 0) or absence (e. g. state 1) of electrons in a specific location (the charge storage node). In order to prevent losses of the stored charge, the storage node is defined by energy barriers of sufficient height E_b to retain charge (as shown in Fig. 5).

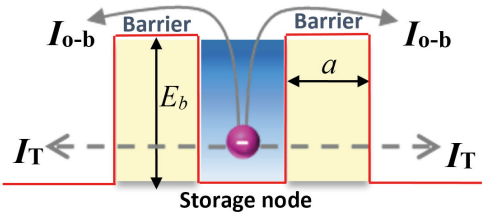


Fig. 5: *A generic electron charge-based memory element*

As discussed above, there are two fundamental mechanisms for the losses of the stored charge. The first is the thermal over-barrier transitions (thermionic emission), which is related to the Boltzmann probability (5). The electron escape frequency is given by:

$$f_{therm} = f_0 \exp\left(-\frac{E_b}{k_B T}\right) \tag{12a}$$

Where $f_0 \sim 10^{12}$ s^{-1}nm^{-2} is the thermal attempt frequency.

Correspondingly, the retention time for one electron in a system with cross-sectional dimension L is

$$t_r = \frac{1}{L^2 f_{therm}} = \frac{1}{L^2 f_0} \exp\left(\frac{E_b}{k_B T}\right) \tag{12b}$$

For a specified t_r, the required minimum barrier height is:

$$E_{b_{\min}} = k_B T \ln\left(t_r \cdot f_0 \cdot L^2\right) \tag{12c}$$

In the case of the 'nonvolatility requirement', i.e. t_r>10 years, gives $E_{bmin} \geq 1.42$ eV at T=300 K. For small retention times, e.g. 50-100ms, typical for DRAM, (1c) yields $E_{bmin} \geq 0.8$ eV.

A second source of charge loss is electron tunnelling. The tunnelling escape frequency for a rectangular barrier is:

$$f_T = f_0^* \cdot \exp\left(-\frac{2\sqrt{2m}}{\hbar} \cdot a \cdot \sqrt{E_b}\right) \tag{13a}$$

Where $f_0^* \sim \sim 10^{13}$ s^{-1}nm^{-2} is the tunnelling attempt frequency.

The electron escape time due to tunnelling is:

$$t_T = \frac{1}{L^2 f_0^*} \exp\left(\frac{2\sqrt{2m}}{\hbar} \cdot a \cdot \sqrt{E_b}\right) \tag{13b}$$

Suppose that the barrier height is large enough to suppress over-barrier escape, i.e. $E_b >> E_{b\min}$, where $E_{b\min}$ is given by (12c). In this case, the store time will be determined by the tunnelling time, t_T: $t_s \approx t_T$. The minimum barrier width for a specified store time, can be estimated from (13b), e.g. for t_s=10 years:

$$a_{\min} = \frac{\hbar}{2\sqrt{2mE_b}} \ln\left(f_0^* \cdot t_s \cdot L^2\right) \tag{13c}$$

As a numerical estimate for t_s>10 years, $E_{b\min} \geq$ 1.42 eV, $m=m_e=9.11\times10^{-31}$ kg, and T=300K, (13c) gives $a_{min} \sim$ 5 nm.

As follows from the above, in order to obtain an electronic memory cell, sufficiently high barriers must be created to retain the charge for a long period of time. In physical systems the barrier can be created by combining materials with different properties, such as e.g. conductor-insulator layered structure used in flash memory (Fig. 6). In a flash memory cell, the charge is stored in a conductive electrode surrounded by insulators (floating gate). In order to prevent loss of the stored charge, the storage node is defined by energy barriers of sufficient height E_b to retain charge. Such barriers are formed by using layers of insulator (I), which surround a metallic storage node (M). The barrier height E_b is fixed, as it is a material-specific property. A simple estimate (12c) for the minimal barrier height and width to satisfy the 'nonvolatility requirement' yielded $E_{bmin} \geq 1.42$ eV and a minimum width of ~5 nm. More detailed calculations in [5] that take into account different practical constrains result in the somewhat larger barrier height of 1.73 eV required to achieve 10 y retention. Such barrier height can be achieved only with a limited number of materials, some of which are listed in Table 1. Note that the effective electron mass in solids is, in most cases, smaller than the *free electron mass* (used in the simple estimate above). According to (13c) the smaller mass will result in a wider barrier or thicker insulator layer (Table 1). The theoretical barrier width must be >5nm for all known dielectric materials (typically >7nm in practical devices). The corresponding practical minimum size of the floating gate cell is ~10 nm [5].

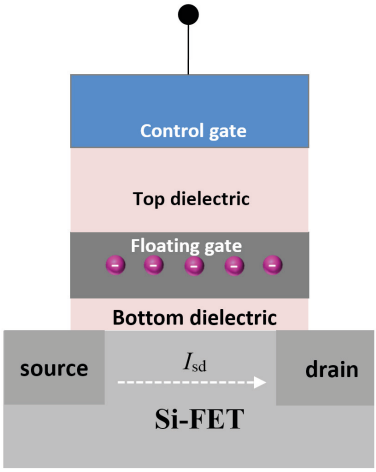


Fig. 6:
Flash memory

Material	Dielectric constant, K	Barrier height, E_b (with Si)	Effective electron mass, m^*	a_{min}
SiO_2	3.9	3.1 eV	$0.50m_0$	5.0 nm
Si_3N_4	7.6	2.4 eV	$0.43m_0$	6.0 nm
Al_2O_3	9	2.8 eV	$0.30m_0$	6.8 nm

Table 1: *Insulator material parameters and the corresponding theoretical minimum insulator thickness for floating gate nonvolatile storage.*

3.4 Spintronic ICT

In addition to charge, electrons possess intrinsic angular momentum (spin). As result, they also possess a permanent magnetic moment [6]:

$$\mu_s = \pm\frac{1}{2} g \cdot \mu_B \tag{14}$$

μ_B is the Bohr magneton, $\mu_B = \frac{e\hbar}{2m_e} = 9.27 \cdot 10^{-24}\ \frac{J}{T}$

g is the coupling constant known as the Landé gyromagnetic factor or g-factor. For free electrons and electrons in isolated atoms g_0=2.00. In solids, consisting of a large number of atoms the effective g-factor can be different from g_0.

The energy of interaction, $E_{\mu-B}$, between a magnetic moment, $\vec{\mu}$, and a magnetic field, $\vec{B}$ is:

$$E_{\mu-B} = -\vec{\mu} \cdot \vec{B} \tag{15}$$

For the electron spin magnetic moment in a magnetic field applied in the z direction, the energy of interaction takes two values depending of whether the electron spin magnetic moment is aligned or anti-aligned with the magnetic field. From (14) and (15) one can write, assuming g=2

$$\begin{aligned} E_{\uparrow\uparrow} &= -\frac{e\hbar}{2m_e} \cdot B_z, \\ E_{\uparrow\downarrow} &= +\frac{e\hbar}{2m_e} \cdot B_z \cdot \end{aligned} \tag{16}$$

The energy difference between the aligned and anti-aligned states represents the energy barrier in the spin binary switch and is

$$E_b = E_{\uparrow\downarrow} - E_{\uparrow\uparrow} = 2\mu_B B_z \tag{17}$$

Equations (43) and (44) represent a physical phenomenon known as Zeeman splitting [6]. The operation of a single spin binary switch is illustrated in Fig. 7. In the absence of an external magnetic field, there is equal probability that the electron has magnetic moment $+\mu_B$ or $-\mu_B$, i. e. the two states are indistinguishable (Fig. 7a). When an external magnetic field is applied (Fig. 7b, c), the two states are separated in energy. The lower energy state has higher probability of population and it represents the binary state '1' (Fig. 7b) or '0' (Fig. 7c) in this system. Binary switching occurs when the external magnetic field changes direction as shown in Fig. 7b and 7c. This abstraction, while very simple, applies to all types of spin devices, including, e.g. spin transfer torque random magnetic memory STT-MRAM [7].

The barrier-forming magnetic field B can be either a built-in field formed by a material layer with a permanent magnetization, or created by an external source.

Let us now consider a hypothetical single spin binary switch that ideally might have atom-scale dimensions. At thermal equilibrium there is a probability of spontaneous transition between spin states '1' and '0' in accordance to (5). Correspondingly, for Π_{err}<0.5, according to (11) the energy separation between to state should be larger than $k_BT\ln 2$.

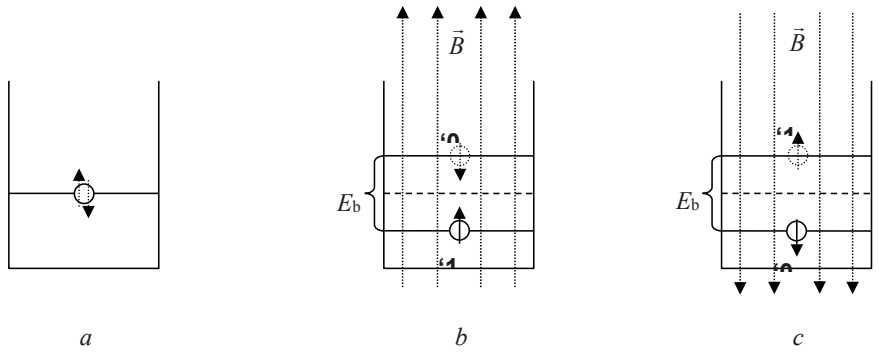


Fig. 7: *An abstract model of a single spin binary switch: (a) B=0, two states are insdistinguishable; (b, c) B≠0, two binary states are separated by energy gap E_b*

$$2\mu_B B_{\min} = k_B T \ln 2 \tag{18}$$

From (18) one can obtain the minimum value of B for a switch operation

$$B_{\min} = \frac{k_B T \ln 2}{2\mu_B} = \frac{m_e}{e\hbar} k_B T \ln 2 \tag{19}$$

At T=300 K Equation (19) results in B_{min}≈155 T. This is much larger than can be practically achieved (examples of technologies to generate magnetic fields are given in Table 2).

Magnet	B	Power	Mass
Small bar magnet	~0.01 T	-	~g
Small neodymium-iron-boron magnet	~0.2 T	-	~g
Big Magnetic-core electromagnet	~2 T	~100 W	~kg
Steady-field superconducting electromagnet [8]	~16 T	~MW	~tons
Non-Destructive Pulsed Magnets at the Dresden High Magnetic Field Laboratory [9]	~90 T	~GW	~tons

Table 2: *Examples of practical implementations of the sources of magnetism.*

Thus, we conclude that single electron spin devices operating at room temperature would require local magnetic fields higher than have been achieved to date with large volume apparatus. In multi-spin systems, it is possible to increase the magnetic moment μ and therefore, to decrease the magnitude of the external magnetic field B required for binary switch operation. The increase of μ can be due to an increase in number of co-aligned spins, which results in collective effects such as *paramagnetism* and *ferromagnetism,* which are considered below.

In a system of N spins in an external magnetic field, there are $N_{\uparrow\uparrow}$ spin magnetic moments parallel to the external magnetic field, and the resulting magnetic moment is

$$\mu = \mu_B \cdot N_{\uparrow\uparrow} = \mu_B N\left(1 - \Pi_{err}\right) = \mu_B N\left(1 - \exp\left(-\frac{\mu_B B}{k_B T}\right)\right) \tag{20}$$

As we saw above for all practical cases $\mu_B B << k_B T$, and since $(1-e^x) \approx x$, for $x \to 0$, there results

$$\mu \approx \frac{N\mu_B^2 B}{k_B T} \tag{21}$$

Equation (21) is known as the Curie law for paramagnetism [6]. From (21) and (18) one can calculate the minimum number of electron spins needed for spin binary switch operating at realistic magnitudes of the magnetic field:

$$N_{\min} \approx \frac{\ln 2}{2}\left(\frac{k_B T}{\mu_B B}\right)^2 \tag{22}$$

For example, for B=0.2 T (small neodymium-iron-boron magnet, see Table 2), $N_{min} \sim 2\times10^6$. If the number of electrons with unpaired spins per atom is f (f varies between 1 and 7 for different atoms), the number of atoms needed is N_{min}/f. Correspondingly, one can estimate the minimum critical dimension a_{min} of the binary switch.

$$a_{\min} \sim \left(\frac{N_{\min}}{f \cdot n_{at}}\right)^{\frac{1}{3}} \tag{23}$$

where n_{at} is the density of atoms in the material structure, e.g. for solid metals $n_V \sim 10^{22}$-10^{23} at/cm^3. Assuming $n_{at} = 1\times10^{23}$ at/cm^3 and B~0.2 T, we obtain a_{min} ~26 nm for f=1 and a_{min} ~14 nm for f=7. Thus we conclude that for reliable operation at moderate magnetic fields, the physical size of multi-spin based devices is larger than the ultimate electron charge-based devices (5-10 nm). The effect of collective spin behavior is currently used in e.g. magnetic random access memory (MRAM, STT-MRAM) [7].

3.5 Scaling below 5 nm

The numbers for smallest characteristic sizes obtained above, i.e. ~5-10 nm for electronic and ~10-20nm for spintronic devices are only 2-3x smaller than those in the leading-edge technology Thus, semiconductor scaling is facing downstream physical limits, which are manifested in (8). This equation also emphasizes factors defining the limits. Since the minimal barrier height E_b is conserved due to error minimization requirements (or equivalently requirements for reduced leakage, enhanced retention, etc.) the mass m of the information-bearing charged particle appears to be the only physical parameter that can be adjusted to continue scaling. In fact, it can be the shown that as device size decreases, there is a corresponding optimal mass value [10]. It should be noted that there is a remarkable discontinuity between the masses of stable particles, from $\sim m_e = 9.11\times10^{-31}$ kg for electrons and $\sim 1800 m_e$ for the protons (e.g. hydrogen atoms). However the use of the atomic masses could still result in a satisfactory device performance, and using atoms as information-bearing particles has obvious advantages for $a < 3$ nm [10].

The suggestion that a heavier mass information carrier may be preferable for nm-scale devices may seem counterintuitive. A common-sense observation is that a lighter mass particle should be easier to move faster and would require less energy. However, this observation is valid only

for constant length transitions. It is easy to show that the switching time at a given energy remains constant for scaled devices as long as the product $L\sqrt{m}$ remains constant [10]. The physical reason, which makes lighter mass less attractive for smaller devices is quantum mechanical tunneling. According to (8), the use of larger mass will decrease tunneling probability and allow continued feature size reduction.

Possible physical realizations for a sub-5 nm binary switch include resistive switches, memristive devices etc., which opens or closes an electrical circuit by the controlled reconfiguration of atoms within an atomic-scale junction. Also, various types of resistive memory based on moving atoms or ions are currently being seriously considered as candidates for ultimately scalable solid-state memory, including phase-change memory (PCM), where rearrangement of atoms between the crystalline and amorphous states results in changes of electrical resistance [11] and 'nanoionic' redox-based resistive memory (ReRAM), where electrochemical effects play an important role [12]. In the following section an expository, physics-based, framework for the estimation of the performance potential and physical scaling limits of ICT devices based on 'moving atoms' is provided. While the emphasis of the below analysis is on nanoionic redox-based memory devices, it is also, to some extent, applicable to other concepts, such as PCM.

4 Scaling limits of nanoionic devices

The term nanoionics implies that electrochemical processes occur in material systems. Examples of nanoionic devices include redox-based memories [12], atomic/ionic switches [13], and memristive devices [14].This section provides analyses of changes in the electrical properties due to addition or removal of a few atoms and stability of a few-atom system.

4.1 Switching mechanisms and the material systems

The operation of nanoioinic devices is based on change in resistance of a metal-insulator-metal (M-I-M) structure. The cell resistor has two stable values: high (OFF), R_{off}, and low (ON), R_{on}. Switching between R_{off} and R_{on} occurs by applying an electrical bias to the resistor.

Mechanisms for resistive switching are based on atomic re-arrangements in a material caused by ion (cation or anion) migration combined with redox processes involving the electrode material or the insulator material, or both [12]. Resistive switching is due to formation of a conductive path within (semi)insulating matrix, and it could be, in the simplest case, that a filament is formed by metal atoms (Fig. 8a), such as growth of Ag dendrites in silver sulfide (Ag_2S) [13]. Another mechanism of resistive switching is modulation of the interface resistance between a metal electrode and the insulating/semiconducting matrix. This can be achieved by re-arrangement of charged defects/impurities near the interface between the semiconductor matrix and an electrode (Fig. 8b). In this case, increased concentration of interface charges change the width and the height of the interface (contact) barrier, and as result the contact resistance changes. The interface resistance mechanism of memory operation was reported for several materials, all of which are metal oxides, e.g. TiO_2 or HfO_2. A filament type mechanism has also been reported to metal oxide systems.

Note that in nanoionic devices, the state is created by moving atoms/ions (WRITE operation), while electron transport is used for sensing the state (READ operation). In the following sections, the mechanisms of both atomic and electronic transport in solids will be considered, and the scaling limits of resistive nanoionic devices are investigated based on the conductance and stability of the nanostructures consisting of a few atoms.

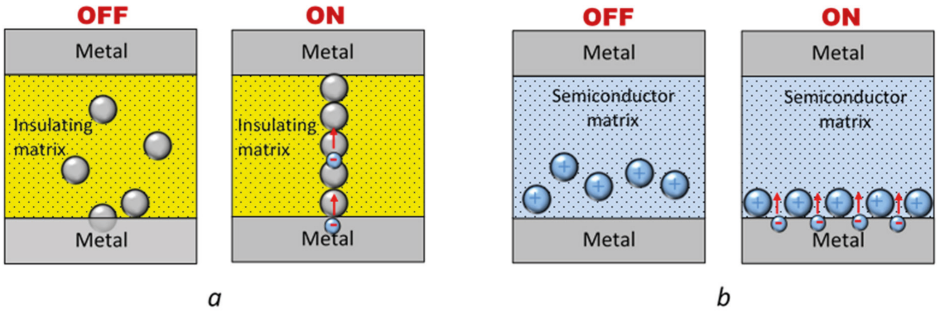


Fig. 8: *Two basic mechanisms of resistance switching due to atomic re-arrangements: (a) conductive bridge (CB) and (b) interface controlled resistance (ICR)*

4.2 Atomic filament: Classical and Quantum resistance

Consider a single metal atom filament bridging two electrodes (Fig. 9). The total resistance of the single-atom filament bridge can be obtained as:

$$R_{bridge} = R_{fil} + 2R_C \tag{24}$$

where R_{fil} is the resistance of the 'filament' and R_C the contact (constriction/spreading) resistance of a contact between the filament and the electrodes. The filament resistance can be calculated as

$$R_{fil} = \rho^* \frac{L}{d^2} \tag{25}$$

where L is the length of the filament, d is the filament diameter (in our case close to the atomic diameter, s), and ρ^* is the resistivity of metal nanowire. The Fuchs-Sondheimer approximation can be used to calculate resistance of nano-scale metal wires:

$$\rho^* = \rho_0 \frac{1-\alpha}{1+\alpha} \frac{\lambda_0}{d} \quad \text{for } d << \lambda_0, \tag{26}$$

where λ_0 is the bulk metal mean free path, derived from the electron concentration in metal using the Sommerfeld model [15, 16, 17], ρ_0 is the bulk resistivity of the metal, and α is specularity, i.e, the probability of an electron being scattered elastically at the side surface of the wire.

The filament diameter, d, is close to the effective diameter of an atom, s, and can be estimated based on the atomic density in a bulk solid metal, n_{at}:

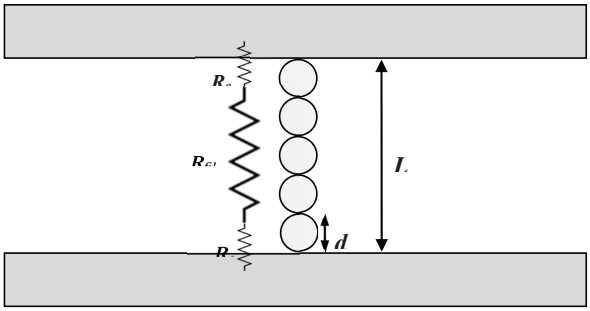


Fig. 9: *Single metal atom bridge*

In general, the values of ρ_0, λ_0, and s are material specific parameters. The corresponding values of these parameters for silver are given in Table 3 (silver is chosen in this paper as an example model material).

Parameter	Numerical value
Atomic density,	n_{at} =5.83·10^{22} cm^{-3}
Effective atomic diameter	s =0.258 nm
Bulk resistivity	ρ_0=15.8 nΩ·m
Electron mean free path	λ_0=54 nm

Table 3: *Parameters of bulk silver used for minimal conductive bridge estimates.*

The length of the filament can be expressed as a function of the number of the constituting metal atoms, N:

$$L = Ns \tag{28}$$

Substituting (26-28) into (25) there results (assuming α=0)

$$R_{fil} = \rho_0 \frac{\lambda_0}{s} \frac{L}{s^2} = \rho_0 \frac{\lambda_0}{s^2} N \tag{29}$$

Next, the constriction (spreading) resistance exists at points of contact between the filament and the electrodes, and is given by [18, 19]:

$$R_C = \frac{\rho_0}{d} \tag{30}$$

Note that in (30) the bulk metal resistivity is used.

Combining (29) and (30) obtain:

$$R_{bridge} = \rho_0 \left(\frac{\lambda_0}{d^2} N + \frac{2}{d} \right) \tag{31a}$$

The smallest possible bridge corresponds to N=1, i.e single-atom contact. The 'classical' resistance (31a) in this case becomes:

$$R_{bridge} = \rho_0 \left(\frac{\lambda_0}{s^2} + \frac{2}{s} \right) \tag{31b}$$

Evaluating (31b) using parameters from Table 3 gives a value for *classical resistance* of the atomic conductive bridge R_{bridge} = 12.94 kΩ.

We now consider *quantum resistance*, which represents the best case conductance through a single atom bridge shown in Fig. 10. It can be estimated as follows. Suppose that a monovalent atom such as Ag, Cu or Au contributes one conductance electron and consider an elementary act of electrical conductance where one electron passes from electrode **A** to electrode **B** with a potential difference between the electrodes of V_{AB} (the corresponding energy change $\Delta E = eV_{AB}$). The speed of the passage process is bounded by the Heisenberg energy-time relation

$$\Delta E \cdot \Delta t \geq \frac{h}{2} \tag{32a}$$

From which the minimum passage time Δt is

$$\Delta t = \frac{h}{2\Delta E} = \frac{h}{2eV} \tag{32b}$$

The current between two electrodes through the atom (involves only one electron at a time) is:

$$I_{AB} = \frac{e}{\Delta t} \tag{33}$$

Putting (32b) into (33), and taking into account Ohm's law, i.e. I=V/R, obtain:

$$I_{AB} = \frac{2e^2}{h} \cdot V = \frac{V}{R_0} \tag{34}$$

where

$$R_0 = \frac{h}{2e^2} = 12.95k\Omega \tag{35a}$$

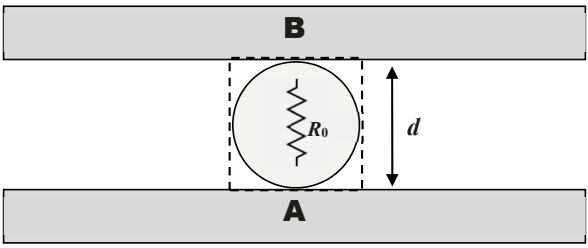


Fig. 10: *Single metal atom bridge*

is *quantum resistance*. A related parameter is quantum conductance:

$$G_0 = \frac{1}{R_0} = \frac{2e^2}{h} \tag{35b}$$

The quantum resistance/conductance sets the limit on electrical conductance in a one-electron channel *in the absence of barriers*.

Experiments with the single atom contacts indeed have demonstrated that the minimum contact resistance is approximately 12.9 kΩ [20]. Therefore, it has been shown that both *classical* and *quantum* models yield a very similar result for the minimal resistance of a single-atom bridge.

4.3 Conductance in the presence of barriers

If a barrier is present in the electron transport system, the conductance will be decreased due to the barrier transmission probability $\Pi_T < 1$. The electrical conductance in the presence of barrier is obtained by multiplying the barrier-less quantum conductance (15b) by the barrier transmission probability Π_T:

$$G = \frac{1}{R} = G_0 \cdot \Pi_T \tag{36}$$

Eq. (36) is a form of the Landauer formula [21] for a one-electron conductive channel.

Suppose that the single-atom bridge in Fig. 10 has the minimum resistance R_0 (i.e. barrierless transport). If now the atom is removed, the two electrodes are separated by the gap of length d, and thus a barrier is formed. The change in the inter-electrode resistance with and without the bridging atom (R_{on} and R_{off}) can be obtained from (36) as

$$\frac{R_{off}}{R_{on}} = \frac{G_0}{G_0 \Pi_T} = \frac{1}{\Pi_T} \tag{37}$$

The Landauer formalism (36) allows for analysis of electron transport in the presence of a barrier in which all different mechanisms of electron transport (conductance) can be expressed through the transmission coefficient Π_T, which is the probability that an electron can transmit through a medium. In the simplest case of one single-electron transmission channel, the electron current is expressed as:

$$I_1 = \frac{2e^2}{h} \cdot V \cdot \Pi_T \tag{38}$$

In the case of transmission through many parallel electron channels, the total current is obtained as a product of I_1 and the number of parallel channels N_{ch}:

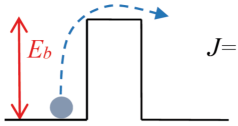
$$I_M = N_{ch} \cdot I_1 = \frac{2e^2}{h} \cdot N_{ch} \cdot V \cdot \Pi_T \tag{39a}$$

which can also be re-written as

$$I_M = N_{ch} \cdot I_1 = \frac{2e^2}{h} \cdot N_{ch} \cdot V \cdot \Pi_T \tag{39b}$$

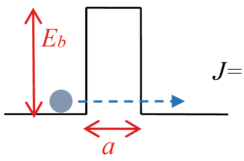
where f_0 is the attempt frequency - is the rate at which electrons available for transmission hit the barrier. For practical cases of multiple-channel transmission the attempt frequency can be calculated based on electron distribution functions in solids, which results in pre-factors given in Table 4.

I. Over-barrier current (thermionic emission)



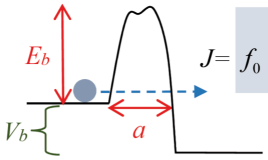
$$J = f_0 = \frac{4\pi e m k_B^2}{h^3} \cdot T^2 \times \Pi_T = \exp\left(-\frac{E_b}{k_B T}\right) \quad (40)$$

IIa. Direct tunneling through simple rectangular barrier ($V \to 0$)



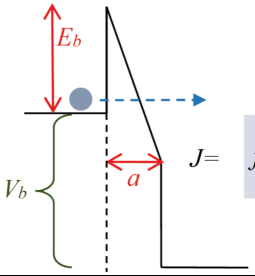
$$J = f_0 = \frac{e^2}{h^2} \cdot \frac{\sqrt{2mE_b}}{a} \cdot V \times \Pi_T = \exp\left(-\frac{2\sqrt{2m}}{\hbar} \cdot a \cdot \sqrt{E_b}\right) \quad (41a)$$

IIb. Direct tunneling through arbitrary barrier ($0 < eV_b < E_b$)



$$J = f_0 = \frac{e^2}{h^2} \cdot \frac{\sqrt{2m}\left\langle\sqrt{E(x)}\right\rangle}{a} \cdot V \times \Pi_T = \exp\left(-\frac{2\sqrt{2m}}{\hbar} \cdot a \cdot \left\langle\sqrt{E(x)}\right\rangle\right) \quad (41b)$$

III. Fowler-Nordheim tunneling through triangle barrier ($eV_b > E_b$)



$$J = f_0 = \frac{e^3}{8\pi h E_b} \cdot \frac{V^2}{a^2} \times \Pi_T = \exp\left(-\frac{2}{3} \cdot \frac{2\sqrt{2m}}{\hbar} \cdot a \cdot \frac{E_b^{\frac{3}{2}}}{eV}\right) \quad (42)$$

Table 4: *Electron transport in the presence of barriers.*

4.4 Barriers in atomic gaps

For a better understanding of the physical origin of the barrier in atomic gaps let us first consider a vacuum-metal interface. Intuitively, a surface barrier must exist to prevent the easy escape of electrons from the metal "reservoir", and thus make it possible for the existence of a stable solid state. A simple (but rather accurate) model for the surface barrier was first proposed by Schottky, based on classical electrostatics [22, 23]. He suggested that electrons leaving a conducting surface and at a given moment located at distance x above the surface must create a positively charged surface layer which attracts the electron (Fig. 11). As result an attractive force appears between the electron and the surface, and thus a barrier exists preventing the

escape. As is shown in the theory of electrostatics, this is equivalent to the force due to a fictitious positive charge located behind the surface at the equal distance (-*x*) as the original charge, i.e. the mirror image of the electron. The attractive force between the electron above the surface and it image below the surface is expressed by the Coulomb law:

$$\mathbf{F} = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{d^2} = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{(2x)^2} = \frac{e^2}{16\pi\varepsilon_0 x} \tag{43}$$

The attractive force acting on the escaping electron is equivalent to a presence of a barrier preventing electron escape. The barrier height, E_b, is equal to the total work to move the electron from a point x near the surface to infinity:

$$E_b = \int_x^\infty \mathbf{F}dx = \int_x^\infty \frac{e^2}{16\varepsilon_0 x^2}dx = -\frac{e^2}{16\varepsilon_0 x} + const = \phi_0 - \frac{e^2}{16\varepsilon_0 x} \tag{44}$$

The integration constant ϕ_0 in (44) is called the work function and is a characteristic property of a given material. The profile of the surface barrier $E_b(x)$ is shown in Fig. 12. At large *x* the barrier can be regarded as rectangular, and the rectangular barrier shape is often assumed in simplified analyses.

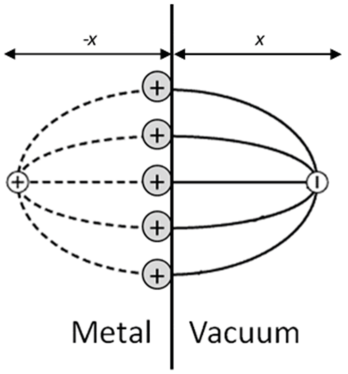


Fig. 11: *Image charge model of barrier formation at metal-vacuum interface.*

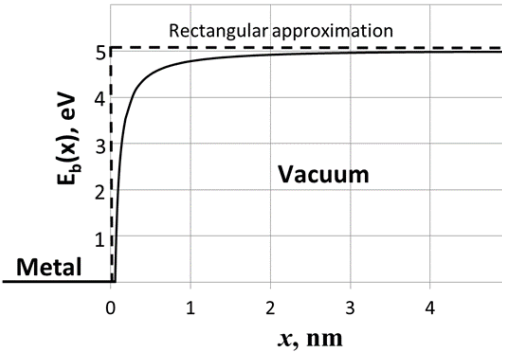


Fig. 12: *Surface barrier profile at metal-vacuum interface.*

Until now we have considered the surface barrier profile, when no external electric fields have been applied to the solid. The effect of an external field *F* is to change the slope of the barrier profile curve corresponding potential energy change of *eFx*:

$$E_b(x) = \phi_0 - \frac{e^2}{16\pi\varepsilon_0 x} - eFx \tag{45}$$

Fig. 13. shows surface barrier profiles for different external electric fields. It can be seen that the barrier height is lower in higher fields, an effect known as Schottky lowering of the surface barrier. This effect can be quantified by taking the derivative of (45) and setting $\frac{dE_b}{dx} = 0$:

$$\frac{dE_b}{dx} = \frac{e^2}{16\pi\varepsilon_0 x^2} - eF = 0 \tag{46}$$

from which the barrier height reduction is

$$\Delta E_b = \phi_0 - E_{b\max} = \sqrt{\frac{eF}{4\pi\varepsilon_0}} \tag{47}$$

and the position of the barrier maximum relative to the surface is

$$x_{\max} = \sqrt{\frac{e}{16\pi\varepsilon_0 F}} \tag{48}$$

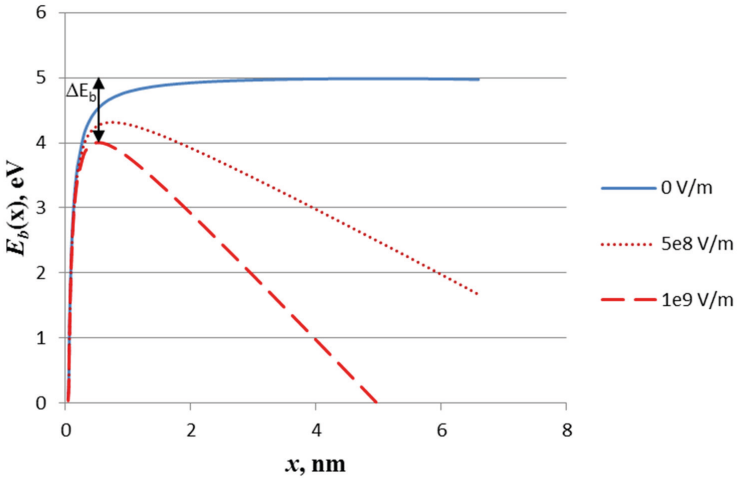


Fig. 13: *Schottky lowering of barrier height due to external fields.*

The above analysis can be extended to the case of two-sided barrier. Consider a vacuum gap of length a between two metal electrodes. There will be an energy barrier formed at the metal-vacuum interface on both electrodes. For larger gaps, if no voltage is applied across the gap, V_{gap}=0, the barrier can be approximated by a rectangular barrier with height equal to the metal work function, ϕ_0 (Fig. 14a). For smaller gaps, the shape of the barrier changes, by reducing barrier height and inducing corner rounding due to image forces as is discussed above. The barrier thus becomes inherently non-rectangular. Extension of Eq. 45 to two-sided barrier results in [24]:

$$E_b = \phi_0 - \frac{e^2}{16\varepsilon_0 K x} - eFx - \frac{e^2}{8\pi\varepsilon_0 K}\sum_{n=1}^{\infty}\left(\frac{na}{(na)^2 - x^2} - \frac{1}{na}\right) \tag{49}$$

where K is the dielectric constant of the material.

The last term in (49) accounts for the interface-to-interface interaction. A useful analytical approximation of (49) was obtained by Simmons [24] in the form:

$$E_b(x) \approx \phi_0 - \frac{e^2 \ln 2}{16\pi\varepsilon_0 K} \cdot \frac{a}{x(a-x)} - \frac{eV_{gap}x}{a} \tag{50}$$

As can be seen in Fig. 14 (a) and (b), for very small gaps, e.g. a <2 nm, the barrier height strongly depends on the gap voltage, V_{gap}. In addition, the effective barrier width for electrons, a_{eff}, is smaller than the inter-electrode 'metallurgical' gap, a. In the following, when a vacuum gap is considered, it is assumed that ϕ_0=4.7 eV (e.g. work function of silver) and K=1, and when the gap is filled by an insulating material, ϕ_0=1 eV and K=7.

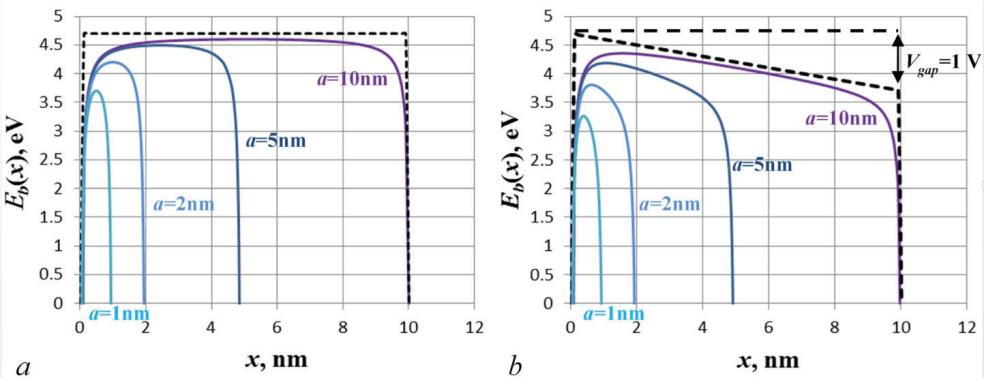


Fig. 14: *Barrier profiles in small Me-vacuum-Me gaps (ϕ_0=4.7 eV, K=1) for three different gap length (1, 2, 5 and 10 nm): a) Unbiased gap, V_{gap}=0, b) Biased gap, V_{gap}=1 Volt*

4.5 Transmission through atomic gaps

Suppose that the single-atom bridge in Fig. 10 has the minimum resistance R_0 (i.e. barrierless transport). If one atom is removed, the two electrodes are separated by an atomic gap, and thus a barrier is formed. If a barrier is present in the electron transport system, the conductance will be decreased due to the barrier transmission probability $\Pi_T < 1$.

The minimum number of atoms needed in a conductive chain to provide sensing margin can now be estimated. This can be done with following thought experiment: the atoms from the chain in Fig. 9 are removed one by one and the resulting change in conductance is calculated. The change in the conductance will be due to barriers formed in such sub-nm gaps as shown in Fig. 15 (calculated using (27b) and (30)).

As shown from numerical values given in the table contained in Fig. 7, a 3-atom gap is sufficient to obtain both a sufficiently large ON current a reasonably large resistance ON/OFF ratio to satisfactorily differentiate the state of a nanoionic device, e.g. the RRAM cell.

N_{at}	a, nm	a_{eff}, nm	E_{bmax}, eV	E_{beff}, eV	I_{on}/I_{off}
1	0.258	0.160	0.38	0.26	2.30
2	0.516	0.426	0.66	0.50	11
3	0.714	0.680	0.75	0.59	73

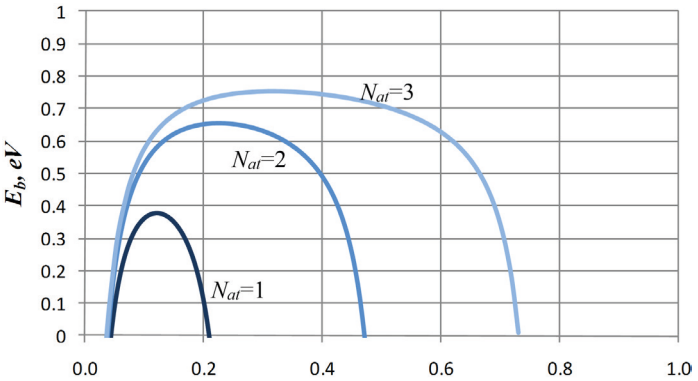


Fig. 15: *Barriers formed as result of removal of 1, 2, and 3 atoms from a single-atom metallic chain embedded in an insulating matrix (ϕ_0=1 eV, K=7). (The corresponding numerical barrier parameters are shown above).*

4.6 Interface Controlled Resistance (ICR)

Another mechanism of resistive switching is re-arrangement of charged defects/impurities near the interface between the matrix and an electrode. In this case, increased concentration of interface charges reduces the width and the height of the interface (contact) barrier, resulting in contact resistance decrease. A simple analysis of this scenario can be performed based on the Mott-Schottky theory of contacts between a metal and a non-metal [25].

The interface resistance mechanism of nanoionic devices was reported for several materials, all of which are metal oxides, such as TiO_2 [26, 27], HfO_2 [28], ZrO_2 [29], ZnO [30] etc. For numerical estimates in this chapter, material parameters of titanium oxide TiO_x are used. It is important to note that metal oxide films are usually non-stoichiometric due to an excess of metal ions or deficiency of oxygen ions [31]. This nonstoichiometry appears to play a key role in the electronic properties of the oxides, since the resulting lattice point defects (e.g. vacancies or interstitial atoms) can electrically act as donors or acceptors. Due to these nonstoichiometric defect levels, the materials often behave as doped semiconductors, and can be described in the framework of a classical semiconductor model. In the past, there were attempts to describe these materials as a special class, called chemiconductors [31, 32] to emphasize three important differences from classical semiconductors: 1) the ions forming donor and acceptor levels in chemiconductors are primarily due to composition variation, 2) the ions can move under electrical fields, while in semiconductors the dopants (i.e. donors and acceptors) don't change their positions, and 3) the distribution of the 'dopants' (donors or acceptors) are inherently non-uniform, especially near the interfaces. Except for these caveats, chemiconductors (i.e. nostoichiometric metal oxides) can be considered as classic semiconductors. For example, the electrical properties of their interfaces with metal electrodes can be described using the Mott-Schottky

model [31, 32] suggesting formation of interface energy barriers (known as Schottky barriers). The Mott-Schottky theory (discussed below) describes the electrical properties of interfaces in terms of the space charge formation in the inerface region, which depends on such macroscopic materials parameters as dielectric constant K and donor concentration N_d. In addition, electron transport across the interface depends on the effective electron mass m^*. The material parameters determining the inteface barrier properties are shown in Table 3.

	HfO_2	TiO_2
Relative static dielectric contant, $\boldsymbol{K}$	14-34	7-114
Band gap, $\boldsymbol{E_g}$	5.1 eV	3.2-3.8 eV
Donor concentartion $\boldsymbol{N_d}$	10^{18} cm^{-3}	10^{20} cm^{-3}
Effective electron mass, $\boldsymbol{m^*}$	$0.15m_0$	$1\ m_0$

Table 5: *Material parameters determining the interface barrier properties.*

When two different materials are brought in contact, an energy barrier is commonly formed at the interface. The origin of the interface energy barrier is in the different concentration and distribution of electrical charges in dissimilar materials. In an extreme case of an interface between a metal (maximum concentration of electrons) and vacuum (zero concentration of electrons), larger barriers are formed often referred to as the work function, ϕ. Now, if two solid materials are brought in contact, they will exchange electrons, and the resulting barrier height can be estimated as the difference between the work functions of the first and the second materials, i.e. $\phi_1 - \phi_2$, called the contact potential difference (measured in volts). In the case of contacts between a metal and a semiconductor or an insulator, the inteface barrier height, Φ, is the difference between the work function of the metal, ϕ_M, and the electron affinity of the semiconductor/insulator, χ_s [25] (if interface states are neglected):

$$\Phi = \phi_M - \chi_s \tag{51}$$

Consider contact between a metal and a *n*-type semiconductor with a concentration of dopants N_d. As a result of the electron exchange, an increased negative charge will accumulate on the metal side of the interface (in an infinitely thin layer for an ideal metal). Due to the charge neutrality requirement, this negative charge must be compensated by an equal positive charge on the semiconductor side (formed by the ionized dopants). Since the concentration of charge carriers in semiconductors is much lower than in metals, the positive charge is formed within some extended layer on the semiconductor side known as a depletion layer of width W. The potential profile in the semiconductor material near the interface can be obtained by solving the Poisson Equation:

$$-\frac{\partial^2 V}{\partial x^2} = \frac{\rho(x)}{\varepsilon_0 K} \tag{52}$$

For the simplest model scenario, the Poisson Equation (52) can be solved assuming uniform distribution of the ionized dopants in the interface layer of width W, and zero net charge outside the interface layer: $\rho(x) = const = eN_d^+$ for $0 < x < W$ and $\rho(x) = 0$ otherwise. We thus can write:

$$-\frac{\partial^2 V}{\partial x^2} = \frac{\rho(x)}{\varepsilon_0 K} = \frac{eN_d^+}{\varepsilon_0 K} \tag{53a}$$

The boundary conditions for integration are: $eV(0)=\Phi$, $V(W)=0$, and $F = \frac{dV}{dx}\Big|_{x=W} = 0$

(zero electric field outside the interface depletion layer). The first integration of (53a) gives the interface electric field distribution $F(x)$:

$$-\frac{\partial V}{\partial x} = -F(x) = \frac{eN_d^+}{\varepsilon_0 K} \cdot x + C_1 \tag{53b}$$

The integration constant C_1 can be found from the boundary condition $F(W) = 0$:

$$\frac{eN_d^+}{\varepsilon_0 K} W + C_1 = 0 \tag{53c}$$

$$C_1 = -\frac{eN_d^+}{\varepsilon_0 K} W \tag{53d}$$

The second integration results in the interface potential profile:

$$C_1 = -\frac{eN_d^+}{\varepsilon_0 K} W \tag{53e}$$

The integration constant C_2 can be found from the boundary condition $eV(0) = \Phi$ which results in

$$C_2 = -\frac{\Phi}{e} \tag{53f}$$

The resulting potential distribution near the interface is

$$V(x) = -\frac{eN_d^+}{\varepsilon_0 K}\left(\frac{x^2}{2} - Wx\right) - \frac{\Phi}{e} \tag{54}$$

The potential distribution $V(x)$ near the interface (54) is plotted in Fig. 16. The zero-bias depletion width W_0 is straightforward to derive from (54) using the condition $V(W)=0$, from which results:

$$W_0 = \sqrt{\frac{2\varepsilon_0 K \Phi}{e^2 N_d^+}} \tag{55a}$$

If an external bias V is applied to the interface, the depletion width of the biased interface from (54):

$$W = \sqrt{\frac{2\varepsilon_0 K(\Phi \pm eV)}{e^2 N_d^+}} \tag{55b}$$

(where plus corresponds to the ‘reverse’ bias and minus to the ‘forward’ bias).

Formulae (34) and (35) represent the parabolic approximation of the interface potential profile, which is most commonly used [25]. For a simple qualitative analysis a linear approximation can also be used [2]. In the linear model, the barrier-forming potential bending is entirely given by the constant interface electric field F, which in turn, depends only on the potential change along distance W (interface depletion width):

$$eF \sim \frac{\Phi \pm eV}{W} \tag{56}$$

(the interface electric field (56) acts both on the electrons and the ionized donors: the negatively charged electrons are repelled from the interface (thus resulting in *depletion*), on the contrary, the positively charged donors are attracted closer to the interface).

As follows from (55b), the depletion width at the M-S interface depends on the concentration of ionized impurities, near interface. The current is then be modulated by changing the impurity concentration and therefore the depletion width at the metal-semiconductor interface.

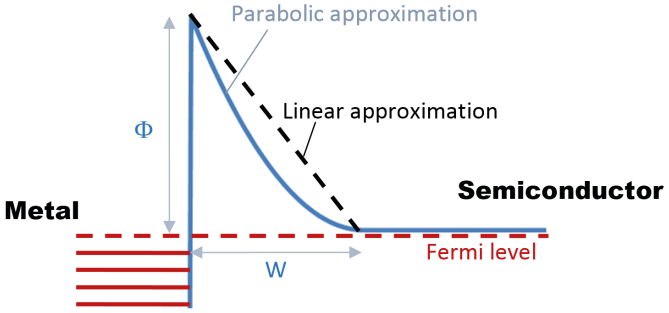


Fig. 16: *Potential distribution near metal-semiconductor interface.*

If finite lateral dimensions of a 3-dimensional semiconductor structure are considered, the side interfaces can also effect the current flow. In simplest model case, these side interfaces are formed between the semiconductor surface and vacuum (SV interface). In practice interfaces with passivation insulator, e.g TiO_2/SiO_2 are representative. Band bending/barrier formation usually occurs at these interfaces, and they need to be taken into account. The band bending results in either depletion (bent up) or accumulation (bent down), and correspondingly, a layer with lower (depletion) or higher (accumulation) conductivity of width W_{SV} is formed as shown in Fig. 17. Therefore, in addition to the depletion WMS layer aligned with the direction of current ('active' interface, modulated by external stimulus), there is a lateral depletion layer W_{SV} perpendicular to the current flow ('passive' interface, which remains more or less stable during device operation). This 'passive' side interface may also effect the total current. If a depletion high-resistive layer of width W_{SV} is formed, the effective cross-sectional area for modulated current flow is decreased. In the case of an accumulation low-resistive layer, a parasitic surface resistor will be formed in parallel with the resistive memory element. In the treatment below, a depletion layer of width W will be considered, which as a typical case of n-type semiconductor. Side depletion effectively reduces the conductive cross-sectional area of the materials system (the blue-colored central region in Fig. 17).

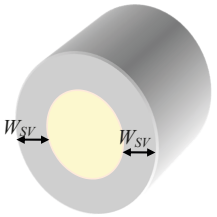


Fig. 17: *Reduction of the effective conduction area due to side depletion*

The barrier height and width determine electron transport through the barrier and thus the contact resistance. In the following, it will be assumed that the contact resistance dominates the total resistance of the structure.

A dominant mechanism of the ohmic conduction through a typical the metal–semiconductor interface is tunneling [2]. For a simple model of a triangular barrier in Fig. 16 (dashed line) the charge flow through the barrier can be calculated using the Fowler-Nordheim (FN) equation (42) (Table 4):

$$J = a\frac{F^2}{\Phi}\exp\left(-b\frac{\Phi^{*\frac{3}{2}}}{F}\right) \tag{57}$$

where a and and b are constants:

$$a_{FN} = \frac{e^2}{8\pi h} \tag{58a}$$

$$a_{FN} = \frac{e^2}{8\pi h} \tag{58b}$$

Note that the barrier height in (57), Φ^*, is reduced compared to Φ due to image force effect in high electric fields: $\Phi^*=\Phi-\Delta\Phi$, where $\Delta\Phi$ is the calculated using (47).

Eq. 57 was used to calculate the interface current as a function of the dopant concentration. A characteristic parameter is interface specific contact resistance:

$$R_c = \frac{V}{J}\left[Ohm\cdot cm\right] \tag{59}$$

Calculations using (59) for parameters of TiO_2 (listed in Table 5) are shown in Fig. 18. To take into account the effect of side depletion in the case of extreme scaling, the current injected from the metal contact through the interface was calculated as

$$I = J\cdot A = J\cdot\left(L-2W_0\right)^2 \tag{60}$$

where A is the effective conductive area of a semiconductor structure with total spatial dimensions L. The effective conduction area is reduced due to side depletion (Fig.17). Fig. 18 shows calculated interface contact resistance with and without the side depletion effects (data for silicon are also shown for reference). As follows from (60), in order to satisfy the condition of constant minimal read current of 1μA, the current density J must increase with decrease of spatial dimensions L. This implies that the interface critical dopant concentration must increase with scaling, as is shown in Fig. 19.

As shown in Fig. 19, for memory cell sizes less than 10 nm, very high concentrations of the interface doping in the order of $N_d > 10^{21}$ cm^{-3} are needed,. The limit on the maximum dopant concentration for the case of TiO_2 can be estimated based on the formation of the Magnéli phases Ti_nO_{2n-1} ($4 \le n \le 10$). A concentration of $N_d=3\times10^{21}$ cm^{-3} corresponds to $n=10$. Therefore the formation of the Magnéli phase $Ti_{10}O_{19}$ can be regarded as the limit for operation of the modulated Schottky barrier interface resistance mechanism. This limit is reached for cell size around L=4 nm.

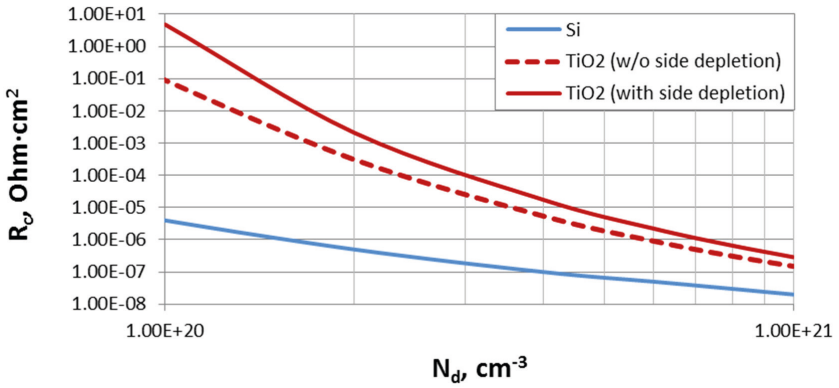


Fig. 18: *Calculated interface contact resistance of a Me-TiO_2 barriers (Φ=0.85 eV).*

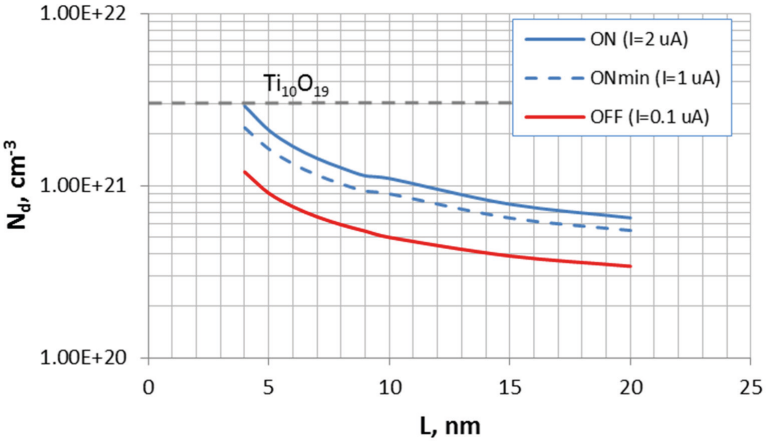


Fig. 19: *Interface critical dopant concentration as a function of the ICR memory cell size.*

Eq. 60 was also used to investigate changes of the interface current as a function of the number of interface dopants. For example, for state retention estimates, the question to be answered is how many dopant atoms Δn need to move from the interface before the ON state is lost? To

address this question, 'operational' parameters were set as I_{ON}~ 2 μA and ON/OFF ratio ~ 20, i.e. twice as much as the minimum read current and minimum ON/OFF ratio specified in the end of Section 2. With such settings, the state is regarded as "lost", when the current through the structure $I < I_{ONmin}$=1μA.

First the ON current was set to be I_{ON}~ 2 μA and the required (critical) concentration of the interface dopants $N_{d,ON}$ was calculated from (53-60). Next, according to the 'benchmark' specifications in the end of Section 2, the resistance ratio was set to be >10, and therefore the OFF current I_{OFF}<0.1I_{ON} ~ 100 nA. The corresponding concentration of the interface dopants N_d OFF was then calculated from (53-60). Finally, the total number of the dopant atoms/defects (or oxygen vacancies in the case of TiO_2) can be calculated for ON and OFF states as follows:

$$n_{on} = N_{d,ON} \cdot L^2 \cdot W_{on} \tag{61a}$$

$$n_{off} = N_{d,OFF} \cdot L^2 \cdot W_{on} \tag{61b}$$

Thus the number of atoms/defects, which need to be moved to/from the interface to enable the specified minimum ON/OFF ratio is

$$n_{off} = N_{d,OFF} \cdot L^2 \cdot W_{on} \tag{61c}$$

The calculation results are summarized in Table 6.

L, nm	N_{on}, cm^{-3} I=2 μA	$N_{on\ min}$, cm^{-3} I=1 μA	N_{off}, cm^{-3} I=0.1 μA	n_{on}	$n_{on\ min}$	n_{off}	$\Delta n_{on\text{-}off}$	$\Delta n_{On\text{-}On\ min}$	W_{on} (1V), nm
20	6.5×10^{20}	5.5×10^{20}	3.4×10^{20}	567	480	296	271	87	2.18
15	7.8×10^{20}	6.5×10^{20}	3.9×10^{20}	349	291	175	174	58	1.99
10	1.1×10^{21}	9.0×10^{20}	5.0×10^{206}	185	151	84	101	34	1.68
5	2.1×10^{21}	1.7×10^{20}	9.0×10^{20}	64	52	42	22	12	1.23
4	2.9×10^{21}	2.2×10^{21}	1.2×10^{21}	49	37	20	29	12	1.05

Table 6: *Scaling-dependent parameters of an ICR cell: the interface dopant concentrations and the corresponding number of dopants which need to be moved to/from the interface to enable the ON/OFF switching.*

4.7 Stability of the minimal nanoionic state

Formation of a conductive bridge (CB) implies alignment of atoms (e.g., metal atoms) between the electrodes to promote electron conduction. At equilibrium, atoms in solids are kept in their position because they are confined in a 'well' between barriers formed by chemical bonds (Fig. 20). The barrier height E_a is often referred to as activation energy (e.g. for diffusion). Dimensions of both the well and the barrier can, to first order, be approximated by the interatomic distance s (27). As a result of thermal excitation, a confined atom vibrates around its equilibrium position with average energy ½ k_BT per degree of freedom and it strikes the barrier with a frequency f_0 (thermal attempt frequency). In some cases when its instantaneous energy exceeds Ea, the atom will jump over the barrier to another site. The rate of such transition f_{tr} is given by the Boltzmann probability:

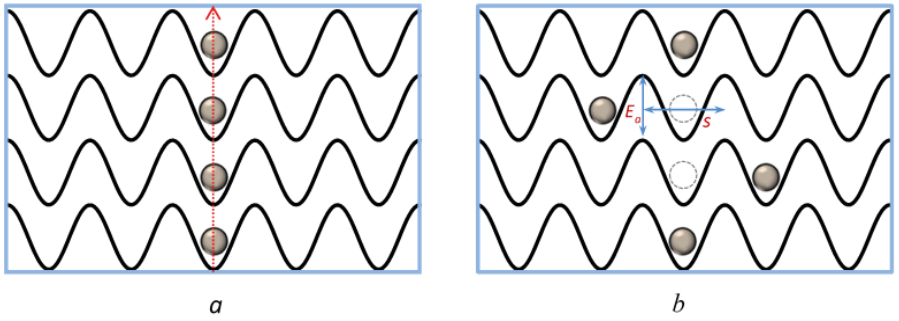


Fig. 20: *A schematic representation of an atomic conductive bridge (CB) depicting barriers for moving atoms within the matrix: a- OFF state; b- ON state*

$$f_{tr} = f_0 \cdot \Pi = f_0 \exp\left(-\frac{E_a}{k_B T}\right) \tag{62}$$

where f_0 is the attempt frequency, determined by an average time τ_0 between barrier strikes, which is determined by atoms's average thermal velocity, u and the interatomic separation s:

$$\tau_0 \sim \frac{s}{u} \tag{63}$$

The velocity u can be found from the kinetic energy relation:

$$\frac{k_B T}{2} = \frac{mu^2}{2} \tag{64}$$

From (63) and (64):

$$\tau_0 \sim s\sqrt{\frac{m_{at}}{k_B T}} \tag{65}$$

Assuming that the movable atoms are of silver (s=0.258 nm, m_{at}=108 a.u.m.=1.79×10^{-25} kg) and T=400K, (65) results in τ_0~1.5 ps (or f_0~7×10^{11} Hz).

The state lifetime Δt can be defined through the probability that n atoms move out from the filament. Let Π be the probability of 'success' in one trial. The number of trials k during time interval Δt is

$$k = \Delta t \cdot f_0 \tag{66}$$

The probability that at least one atom will pass over the barrier E_a during a sampling time Δt is

$$\pi_k = 1 - (1 - \Pi)^k \tag{67a}$$

And the probability for n atoms to escape the filament during the interval Δt is

$$\pi_{kn} = \left(1 - (1 - \Pi)^k\right)^n \tag{67b}$$

Let, π_{kn}=½, then from (67b):

$$\left(1-\exp\left(-\frac{E_a}{k_BT}\right)\right)^{f_0\cdot\Delta t}=1-2^{-\frac{1}{n}} \tag{68a}$$

or

$$\Delta t=\frac{1}{f_0}\frac{\ln\left(1-2^{-\frac{1}{n}}\right)}{\ln\left(1-\exp\left(-\frac{E_a}{k_BT}\right)\right)} \tag{68b}$$

Consider now the resistive switch in ON state (Fig. 20a), with conductive filament formed with a single-atom chain. Let E_a=1 eV, then the lifetime of one atom (n=1) in the filament from (68b) will be about 4s at T=400K and Δt ~ 1800s at T=330K. As was discussed in section 4.5, a 3-atom gap is sufficient to obtain a reasonably large resistance ON/OFF ratio (>10). Repeating calculations for n=3, one obtains Δt~9s at T=400K and Δt~4000s at T=330K. The state lifetime can be increased by adding parallel single-atom chains, to increase the redundancy. For example, if there are two touching parallel conductive chains, the conductance will be broken only if both atoms in the same level jump out of the conductive filament. To further increase redundancy, suppose that minimal conductive bridge forms a 4 atoms × 4 atoms × 4 atoms cube, which is approximately 1 nm in size. According to (68b), such a structure can have a lifetime of ~3×10^8 s (nonvolatile memory applications) if E_a=1.3 eV (T=330K) or E_a=1.56 eV (T=400K). A more detailed stability analysis of atomic filaments can be found in [33].

The above approach can be extended to the stability analysis of interface controlled resistance structures of Fig. 8b. In this case it is convenient to use the time-dependent diffusion equation for a fixed number of ions [34]:

$$N(x,t)=\frac{S_0}{\sqrt{\pi Dt}}\exp\left(-\frac{x^2}{\sqrt{4Dt}}\right) \tag{69a}$$

where D is the diffusion coefficient

$$D=\Delta x^2 f_0\exp\left(-\frac{E_a}{k_BT}\right) \tag{69b}$$

In (69a) $N(x,t)$ is ion concentration and S_0 is the initial 'planar' concentration at the surface. When the initial state is ON, governed by the number of ions n_{on} within the interface depletion width W (see Table 6), then as a simplification, one can assume that initially all of the ions within one atomic jump distance (Δx=s) of the surface reside at the surface, thus $S_0 = n_{on}/L^2$. The solution for the final ion count in the interface region is [33]:

$$n_{final}=n_{on}\cdot\mathrm{Erf}\left(\frac{W}{\sqrt{4D\cdot\Delta t}}\right) \tag{70}$$

In order to estimate the state life time (data retention time), (70) can be solved for Δt assuming the final ion count is known. The ON state is regarded as "lost" when $n_{final} = n_{on\min}$ (see Table 6). For example, for a cell with L of 10 nm, n_{on}=185 and $n_{on\min}$ = 151.

Fig. 20 compares the state lifetimes in CB and ICR models as a function of the critical number of atoms to create the ON state. Both models, i.e. as given by Eq. 68b for CB and Eq. 70 for ICR are in a reasonable agreement. Note that retention decreases for smaller cells.

4.8 Switching speed and energy of ultimate nanoionics devices

Consider the minimal conductive bridge cell in ON state, formed by a 4 atoms × 4 atoms × 4 atoms cube. (64 metal atoms, approximately 1 nm in size). Then, to achieve non-conductive OFF state we partially 'dissolve' the cube by applying external stimulus, such as electrical signal (specific mechanisms can be different for different materials, however they all include moving atoms). As was discussed in section 4.5, a 3-atom gap is sufficient to obtain a reasonably large resistance ON/OFF ratio (>10), therefore it is assumed that 'dissolving' of a 3 atoms × 4 atoms × 4 atoms fragment (48 atoms) represents ON-OFF switching. Suppose that the external stimulus decreases the activation energy E_a in (62) by ΔE for each atom in the bridge cube and enables atomic drift and diffusion in the direction of the applied field. The time to dissolve the bridge can be estimated using (68b):

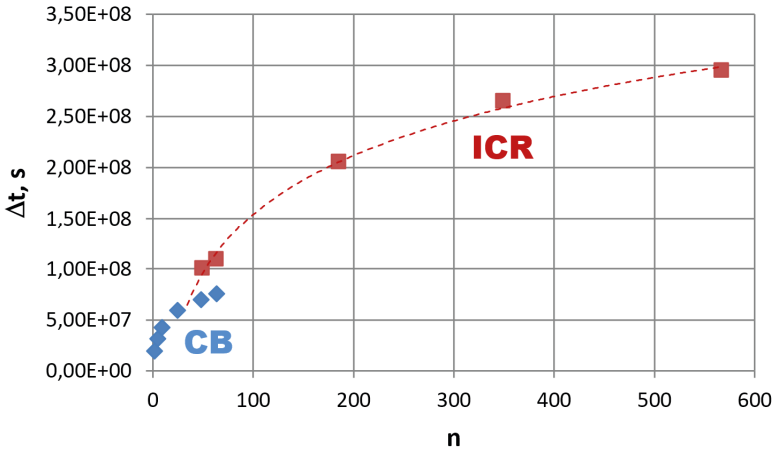


Fig. 20: *State lifetimes in CB and ICR models as a function of the critical number of atoms to create the ON state (E_a=1.25 eV and T=330K)*

$$t_{SW} = \frac{1}{f_0} \frac{\ln\left(1-2^{-\frac{1}{n}}\right)}{\ln\left(1-\exp\left(-\frac{E_a-\Delta E}{k_B T}\right)\right)} \tag{71}$$

For E_a=1.25 eV and ΔE=1eV, $t_{SW} \approx$ 40ns.

Assuming ΔE=1eV=1.6×10^{-19} J, the switching energy can be estimated as

$$E_{SW} \sim 48\Delta E \sim 10^{-17} J \tag{72}$$

It is instructive to compare the result (72) with the energy for fusion (melting) a volume v=1 nm^3 of silver:

$$E_{fusion} = c_{Ag} \cdot m \cdot \left(T_m - T_{op}\right) + H_{fus} \cdot m = \gamma v \cdot \left(c_m (T_m - T_{op}) + H_{fuse}\right) \approx 3.3 \cdot 10^{-18} J \tag{73}$$

(where c_{Ag}=0.233 Jg^{-1}K^{-1} is the specific heat capacity of silver, T_m=1235 K is the melting point of silver, T_{op}=330 K is the device operating temperature, H_{fus}=104.4 J/g is the latent heat of fusion of silver, and γ=10.49 g/cm^3 is the density of silver).

Note that (72) represents an estimates of a lower bound of switching energy of an nanoionic device, as it assumes a 100% conversion efficiency of the external stimulus into the activation barrier lowering ΔE, which be difficult to achieve in practical devices. If, for example, the conversion efficiency is 10%, the resulting switching energy is E_{SW}~10^{-16} J.

5 Summary

Based on the idea that information is represented by the state of a physical system, e.g., the location of a particle, we have shown that energy barriers play a fundamental role in evaluating the operating limits of information processing devices. In order for the barrier to be useful in information processing applications, it must prevent changes in the state of the processing element with high probability, and it also must support rapid changes of state when an external command is given. If one looks at the limit of tolerable operation, that is, the point at which the state of the information processing element loses its ability to sustain a given state, it is possible to advance estimates of the limits of scaling and performance for various kinds of information processing elements.

A generic physics-based abstraction for ICT devices was introduced and applied to several ICT technologies. The scaling limits of electron-based devices are~5-10 nm due to quantum-mechanical tunneling. Smaller devices can be made, if information-bearing particles are used whose mass is greater than the mass of an electron. Therefore the new principles for devices, scalable to ~1 nm, could be 'moving atoms' instead of 'moving electrons'.

Theoretical feasibility of the 1-nm devices was justified based on electrical properties of the few-atom systems.

References

[1] R. U. Ayres, Information, Entropy and Progress, AIP Press, New York, 1994..

[2] H. Schroeder, V. V. Zhirnov, R. K. Cavin, R. Waser, "Voltage-time dilemma of pure electronic mechanisms in resistive switching memory cells", J. Appl. Phys. 107 (2010) 054517.

[3] Gomer, R. Field Emission and Field Ionization. s.l.: Harvard University Press, 1961.

[4] C. Cohen-Tannoudji, B. Diu and F. Laloë. Quantum Mechanics (Hermann and John Wiley & Sons, 1977).

[5] V. Zhirnov and T. Mikolajick, Chapter 26: Flash Memories, in: Nanoelectronics and Information Technology, by Rainer Waser (Ed.) (Wiley 2012).

[6] J. Singh, Quantum Mechanics - Fundamentals and Applications to Technology (John Wiley & Sons, 1997).

[7] V. Zhirnov and T. Mikolajick, Chapter 26: "Flash Memories", in: Nanoelectronics and Information Technology, by Rainer Waser (Ed.) (Wiley 2012).

[8] A. F. Lietzke, S. E. Bartlett, P. Bish, S. Caspi, D. Dietrich, P. Ferracin, S. A. Gourlay, A. R. Hafalia, C. R. Hannaford, H. Higley, W. Lau, N. Liggins, S. Mattafirri, M. Nyman, G. Sabbi, R. Scanlan, and J. Swan-son, "Test results of HD1b, and upgraded 16 Tesla Nb_3Sn Dipole Magnet", IEEE Trans. Appl. Supercond. 15 (2005) 1123.

[9] S. Zherlitsyn, T. Herrmannsdörfer, B. Wustmann, and J. Wosnitza, "Design and performance of non-destructive pulsed magnets at the Dresden High Magnetic Field Laboratory", IEEE Trans. Appl. Supercond. 20 (2010) 672.

[10] V. V. Zhirnov and R. K. Cavin, "Emerging research nanoelectronic devices: The choice of information carrier", ECS Transactions 11(6) (2007) 17.

[11] R. Jeyasingh, E. C. Ahn, S. Burc Eryilmaz, S. Fong, H-S. P. Wong, Chapter 5: "Phase Change Memory", in: Emerging Nanoelectronic Devices, by A. Chen, J. A. Hutchby, V. V. Zhirnov, G. I. Bourianoff (Eds) (Wiley 2014).

[12] S. Menzel, E. Linn, R. Waser, Chapter 8: "Redox-based Resistive Memory", in: Emerging Nanoelectronic Devices, by A. Chen, J. A. Hutchby, V. V. Zhirnov, G. I. Bourianoff (Eds) (Wiley 2014).

[13] T. Hasegawa, K. Terabe, T. Sakamoto, M. Aono, "Nanoionics Switching Devices: Atomic Switches" MRS Bull. 34 (2009) 929.

[14] D. B. Strukov and R. S. Williams, "Exponential ionic drift: fast switching and low volatility of thin-film memristors", Appl. Phys A 94 (2009) 515.

[15] F. J. Blatt, Physics of Electron Conduction in Solids (McGraw-Hill New York 1968).

[16] R. B. Dingle, "The electrical conductivity of thin wires", Proc. Roy Soc. London, 47 (1950) 545.

[17] E. H. Sondheimer, "The mean free path of electrons in metals", Adv. Phys. 1 (1952) 1-42.

[18] A. G. M. Jansen et al., "Point-contact spectroscopy of metals", J. Phys. C: Solid St. Phys. 13 (1980) 6073

[19] Norberg G. et al. “Contact resistance of thin metal film contacts”, IEEE Trans. Comp. Pack. Technol. 29 (2006) 371.

[20] E. Scheer, N. Agrait, J. C. Cuevas, A. L. Yeyati, B. Ludoph, A. Martin-Rodero, G. R. Bollinger, J. M. van Ruit-enbeek, C. Urbina, “The signature of chemical valence in the electrical conduction through a single-atom contact”, Nature 394 (1998) 154.

[21] Y. Imry and R. Landauer, “Conductance viewed as transmission”, Rev. Mod. Phys. 71 (1999) S306.

[22] W. Schottky, “The influence of the structural effects, especially the Thomson graphic quality, on the electron emission of metals”, Phys. Zs. 15, (1914) 872.

[23] R.O. Jenkins and W.G. Trodden, Electron and Ion Emission from Solids, (Dover Publications, Inc, New York, 1965).

[24] J. G. Simmons, “Potential barriers and emission-limited current flow between closely spaced parallel metal electrodes”, J. Appl. Phys. 35 (1964) 2472.

[25] S. M. Sze, Physics of Semiconductor Devices (John Wiley & Sons, Inc, 1981).

[26] J. J. Yang, M. D. Pickett, X. M. Li, D. A. A. Ohlberg, D. R. Stewart, R. S. Williams, “Memristive switching mechanism for metal/oxide/metal nanodevices”, Nature Nanotechnology 3 (2008) 429.

[27] D. S. Jeong, H. Schroeder, R. Waser, “Mechanism for bipolar switching in a Pt/TiO2/Pt resistive switching cell”, Phys. Rev. B 79 (2009) 195317.

[28] P. Gonon et al. “Resistance switching in HfO_2 metal-insulator-metal devices”, J. Appl. Phys 107 (2010) 074507.

[29] B. Sun et al., “The Effect of Current Compliance on the Resistive Switching Behaviors in TiN/ZrO2/Pt Memory Device”, J. Appl. Phys. 48 (2009) 04C061.

[30] L. M. Kukreja, A. K. Das, P. Misra, “Studies on nonvolatile resistance memory switching in ZnO thin films”, Bull. Mat. Sci. 32 (2009) 247.

[31] J. W. Scultze and M. M. Lohrengel, “Stability, reactivity and breakdown of passive films. Problems of recent and future research”, Electrochimica Acta 45 (2000) 2499.

[32] U. Stimming and J. W. Schultze, “A semiconductor model of the passive layer on iron electrodes and its application to electrochemical reactions”, Electrochimica Acta 45 (1979) 859.

[33] V. V. Zhirnov, R. Meade, R. K. Cavin, and G. Sandhu, “Scaling limits of resistive memories”, Nanotechnology 22 (2011) 254027.

[34] W.R. Runyan, K.E. Bean, Semiconductor Integrated Circuit Processing Technology (Addison-Wesley Publishing Company, 1990), P.390

E 3 Select Devices For Memristive Crossbar Arrays

Dirk J. Wouters
Institute of Electronic Materials, IWE-2
RWTH Aachen University

Contents

1 Introduction: need for select devices 2

2 General principles and different types of select devices 5
2.1 Rectification and/or non-linearity ? 5
2.2 Types of select devices 6
2.3 Alternatives to select devices 7
2.4 Selection and current compliance 9

3 Target specifications for select devices 9
3.1 General *I-V* requirements. 9
3.2 Specific requirements based on circuit analysis 10
3.3 Operation speed 11
3.4 Reliability and variability 11
3.5 Scalability 12

4 Material, integration, and scaling constraints 12
4.1 2D, stacked 2D, and 3D memory arrays 12
4.2 Integration and scaling issues of Si-based select devices 13
4.3 Metal-oxide (MO_x) versus classical semiconductor devices 14
4.4 Other materials for selector devices 14

5 Overview of proposed select devices 15
5.1 Select Transistors 15
5.2 Diodes 15
5.3 Bipolar two-terminal selectors: 17
5.4 Self-Selecting Devices 22

6 Conclusions 24

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction: need for select devices

Memristive devices eventually have to be integrated into large arrays. The two-terminal nature of the device, together with the extreme device scalability (down to a few 10's of nm^2 [1]), in principle enables the fabrication of very dense two-dimensional (2D) crossbar arrays. "Raw" crossbars, with a single memristive memory device located at each crossing of bitline (BL) and wordline (WL), form the simplest possible and highest density array configuration (see Fig. 1), and have been proposed earlier e.g. for magnetic core based memories [2].

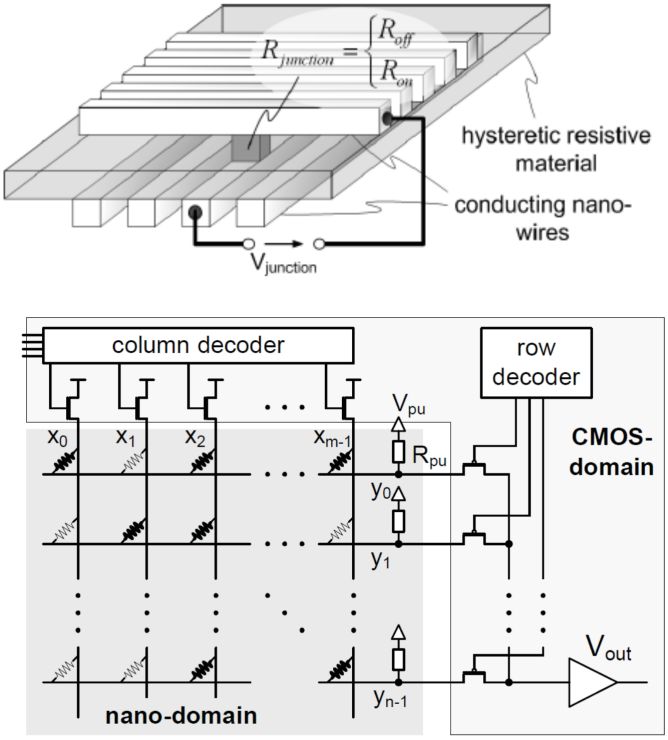


Fig. 1: *Crossbar configuration : (a) raw array [3]; (b) crossbar with external CMOS circuitry for row and column selection as well as readout sensing[4]*

Selection of a certain memory element, either for writing or reading it, is achieved by proper biasing of BL's and WL's, and different bias schemes as *V*/2 and *V*/3 have been proposed in order to optimize the array performance, see Fig. 2.

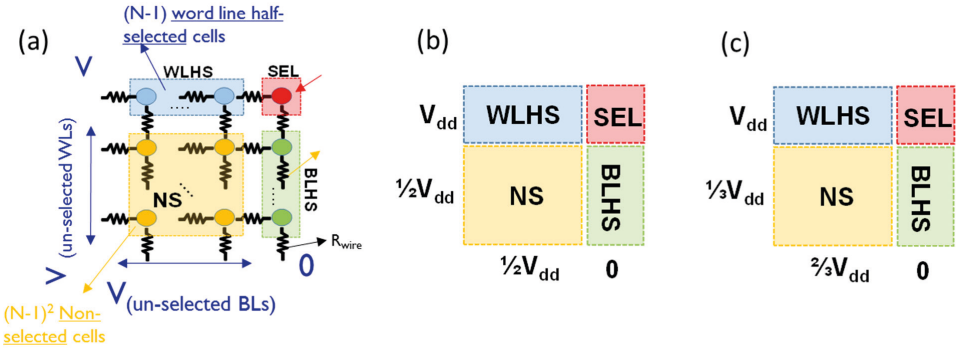


Fig. 2: *Selection of a particular memory cell inside a crossbar array: (a) schematic; (b) V/2 scheme; (c) V/3 scheme. (SEL= selected cell, NS= non-selected cells, WLHS= word line half selected cells, BLHS= bit line half selected cells). The figure is drawn for worst-case selected cell, i.e. largest resistive voltage drop over word and bit line[5]*

However, even so, raw crossbar arrays of memristive devices are still prone to both read and write errors, as well as large power consumption during read:

(i) read errors: as memristive devices have (nearly) linear resistive *I-V* characteristics in their different states, sneak-path currents flowing through half or non-selected devices can increase the apparent device current, masking a high resistive state (Fig. 3);

(ii) write disturb: the limited write time non-linearity with voltage allows for (slow) write even at reduced voltages (cf. voltage-time dilemma[6]) and hence may result in unwanted write of the half and non-selected cells (especially of the half-selected cells in the *V*/2 regime);

(iii) large power consumption: currents through half (*V*/2 scheme) and/or non-selected cells (*V*/3 scheme) can, by their number, result in large power consumption especially during write.

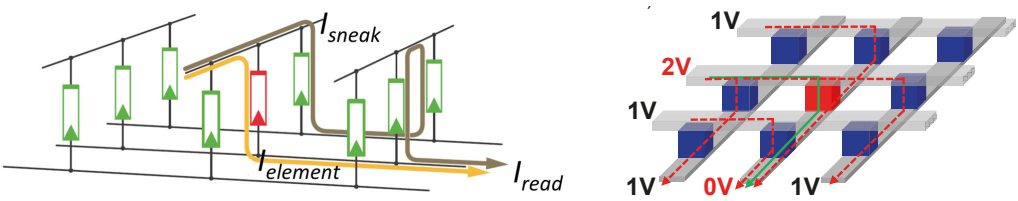


Fig. 3: *Illustration of parasitic currents in cross-point array with (nearly) linear behaved resistive memories: (a) sneak currents during read[7];(b) currents flowing in WLHS and BLHS cells during write operation (V/2 scheme example)[5]*

These issues strongly limit the implementation of memristive memory devices in raw crossbars except for ultra-small arrays [8] (more detailed analysis of the issues of raw crosspoint arrays are found in [4] and [9]). Merely increasing the ON/OFF resistance ratio of the memristive element does not solve these issues, on the contrary, a detailed analysis indicates there is an optimal high resistance state resistance value above which the read margin decreases[10].

A solution of the above problems, however, is the use of a select device (selector) in combination with the memristive device in each memory cell (Fig. 4).

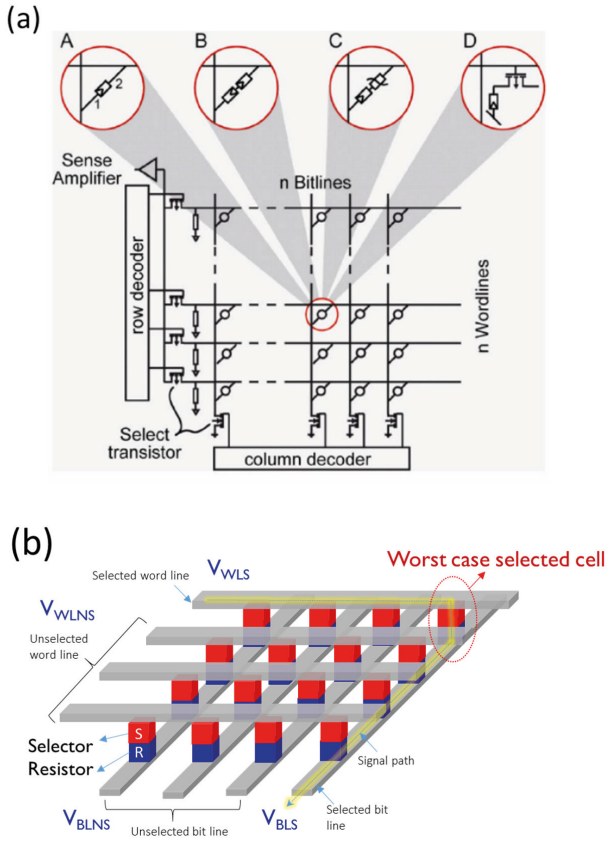


Fig. 4: *(a) Crossbar array of memory cells existing of A:single memristive element; B:2 antiparallel memristive elements (complementary switching cells); C: memristive element and two-terminal selector element, D: memristive element with transistor selector [11]; (b) Dense crossbar array using vertical stacked selector and memristive element (1S1R cell) [5]*

Using a two-terminal select device that is vertically stacked with the memristive device still results in a similar small cell area (see Fig. 4b). This chapter will describe the requirements for such select devices as well as give an overview of the different types of select devices proposed. An interesting recent review article can be found in [12].

2 General principles and different types of select devices

2.1 Rectification and/or non-linearity ?

The select device should avoid the read and program currents mentioned above as well reduce the power consumption. For that, it should prevent the occurrence of sneak-current paths and program disturbs, as well as reduce the current through partially biased cells. Two basic principles can be used to attain such selectivity: rectification and non-linearity (or combinations hereof), see Fig. 5 :

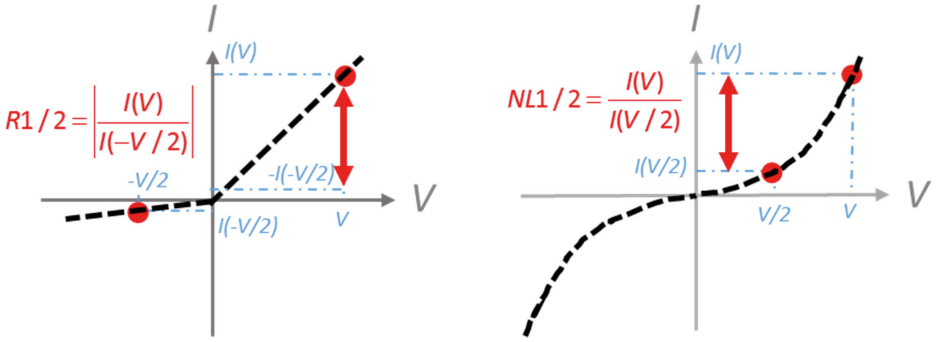


Fig. 5: *Basic principles for selection: (a) rectification, (b) non-linearity*

(i) rectification (Fig. 5a): implies the conduction of the current for only one voltage polarity over the device. It is clear that this can prevent read errors as it will block part of the sneak current path in Fig. 3a. Also, while it will not reduce the leakage current of half selected cells (*V*/2 scheme), it will reduce the leakage current of the non-selected cells (in *V*/3 scheme). However, it will not prevent the write disturb errors of half selected cells.

(ii) non-linearity (Fig. 5b): a strong nonlinear *I-V* characteristic in the ON state (or low resistive state LRS) will limit the current conduction at low voltages, while allowing full current transmission at high voltages. This will prevent sneak-current path induced leakage current while it also suppresses write disturb (for an analysis of the influence of cell non-linearity on the array performance, see e.g. [13] and [14]). A widely used characteristic quantifying the device non-linearity is the half bias non-linearity *NL*1/2 defined in (1), with *I(V)* the current through the selector at voltage *V* over the selector, and V_{op} the highest voltage over the selector, i.e. during write :

$$NL1/2 = \frac{I(V_{op})}{I(V_{op}/2)} \tag{1}$$

It is clear, however, that a rectifying device cannot be used for bipolar switching memristive devices that need (more or less symmetric) opposite current polarities for the SET and RESET operation. In that case, a nonlinear device with symmetric non-linear I-V characteristics is required.

2.2 Types of select devices

We make here a distinction between four major types of select devices:

Select Transistors

The ideal select device is the field-effect transistor, combining with the memristive element in a so-called 1T1R cell structure. By control of the gate voltage, excellent non-linearity of the drain-source current can be obtained. The device also operates bidirectional, although its symmetry is somewhat compromised by the body effect (i.e., the dependence of the threshold voltage on the bulk to source bias). The Si-based Metal-Oxide-Si Field Effect Transistor (MOSFET) is also a very well established device and standardly available in CMOS technologies. However, its use is limited to 2D planar arrays (with the transistor fabricated in the substrate). Also, the cell structure is quite large as the standard transistor is a planar device with 3-terminals and the occupied transistor area further scales with the required program current and operation voltages. A solution to that could be vertical transistor structures that can be stacked with the memristive element to achieve minimal cell size, for which e.g. CMOS process compatible vertical bipolar transistors have been proposed [15]. 3D layer stacking is also possible using different materials from single-crystal Si (polycrystalline Si [16]or even semiconducting metal oxides [17]).

Diodes

The simplest two-terminal select device is a diode. A diode (either a junction diode or a Schottky barrier diode) combines rectification (by blocking reverse current) with a strongly non-linear forward *I-V* characteristic. By this, it forms a good select device, however only for unipolar switching memristive devices (e.g. for phase change memory). As a two-terminal device, it can be stacked with the memristive element to achieve a high density 1D1R cell. Using polycrystalline Si [18] or metal-oxide based diodes (see e.g.[19]), also vertical stacking of arrays is possible.

Bipolar two-terminal selectors: Type I: Symmetric Diodes (Fig. 6a,c)

A select device with symmetric bipolar non-linear *I-V* characteristic which is continuous and without a negative differential resistance (NDR) region, is typically called a symmetric diode. This is one of the two main types of two-terminal selector devices for realizing a 1S1R cell. Such device characteristic can be realized based on different physical concepts (e.g. semiconductor devices as punch-through diodes, back to back Schottky diodes (also called Baritt diodes), tunnel barrier devices, etc.), and different materials (Si, metal oxides,…).

Bipolar two-terminal selectors: Type II: Threshold Switches (Fig. 6b,d)

A threshold switch device has a bipolar non-linear *I-V* characteristic showing negative differential resistance (“ON-switching”) when biased beyond a certain threshold voltage. Often, the characteristic is also discontinuous, with a different return characteristic and “OFF-switching” at a certain holding current or voltage. Again, different physical mechanisms can cause threshold switching as electronic switching in phase change materials (Ovonic switch [22]) and insulator-to-metal transitions (IMT) in metal oxides [23]. This is the second main type of two-terminal selector devices for 1S1R cells.

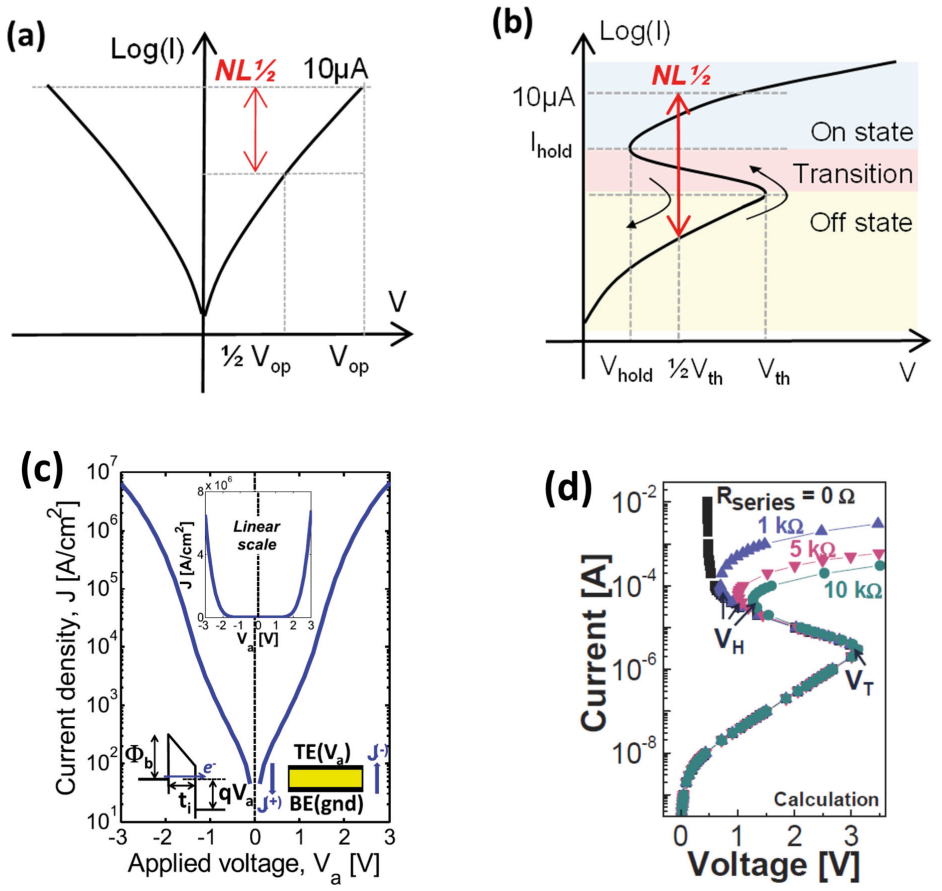


Fig. 6: *Schematic (a,b) and prototypical (c,d) I-V characteristics of two-terminal symmetrical selectors of Type I "Symmetric Diode"(a,c) and Type II "Threshold Switching Device" (b,d) (a,b : [5], c:[20], d:[21])*

2.3 Alternatives to select devices

Self-rectifying/self-selecting devices

Instead of adding a separate selector device, one could think of building rectification and/or non-linearity into the memristive memory device itself. This would in particular be interesting for building vertical three-dimensional (3D) memory arrays (see section **4.1**) where it is not possible to stack a selector on top of a memory device (as, e.g., the intermediate electrode would short a full column of memory cells). Such devices initially have been called self-rectifying cells (SRC), however, because non-linearity is a more desired property especially for bipolar switching device, the more general name of self-selecting device (SSD) is more appropriate.

Different SSD device concepts have been proposed, many containing bi- or multilayers of metaloxides (and often including TiO_x), see section **5.4**. Different from standard oxide ReRAM devices, their switching characteristics are also not always filamentary but in some cases claimed to be areal switching.

Complementary resistive switching (CRS) cells

An interesting alternative cell structure is the anti-serial stacking of two resistive switching cells, in a so-called complementary resistive switching configuration [7] (Fig. 7). The combined *I-V* characteristics show strongly non-linear behavior allowing for a good cell selection. The advantage of this CRS device is that no new technology is required for the selectors. The main drawback is that the normal,(current) read-out is destructive, although an alternative non-destructive capacitance based readout has been proposed[24].

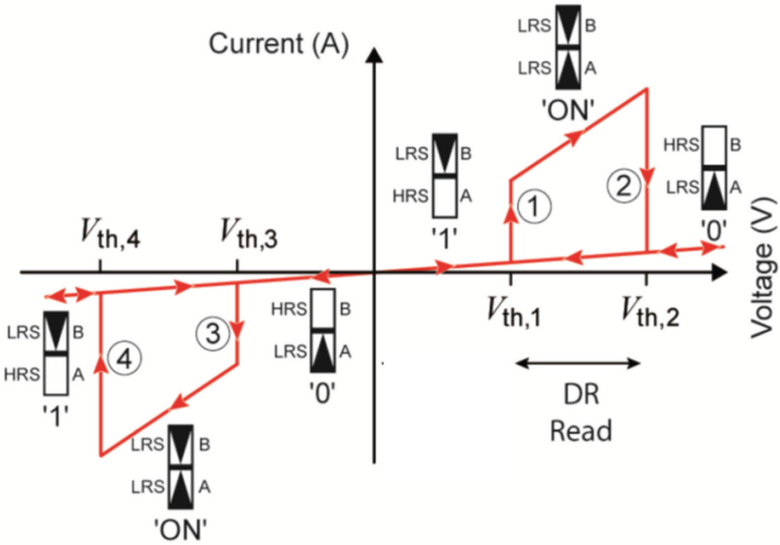


Fig. 7: *Switching characteristics of a complementary resistive switching (CRS) cell, composed of 2 back-to back memristive elements (A and B). The different resistive states of the two cells are illustrated (ON state is depicted by black triangular filament) [7]*

CRS cells based on both metal cation based Conductive Bridging RAM (CBRAM, also called electrochemical memory (ECM) [25]) and metal oxide ReRAM (valence change memory devices (VCM) [25]) have been proposed, see e.g.[26], [27], [28], and[29]. Also, it has been observed that complementary switching can be obtained in a single cell [30], called Complementary Switching (CS) cell, by optimization of the cell stack structure and operation conditions. Note that CRS and CS cells operate in self-compliant current mode (see section **2.4** below).

2.4 Selection and current compliance

Programming of memristive memory devices (at least of those that switch filamentary) requires the application of a controlled current limitation during both SET and FORMING [31]. This requires a short, low-parasitic capacitance and low-resistive connection path between the current-controlling device and the memristive element [32]. Again, the ideal memory cell is the 1T1R cell, where the transistor can act both as the current control device as well as the selector device. For the other types of selector devices, current control is not included in the cell and has to be achieved by transistors driving the WL (or BL). Line resistance and parasitic capacitances may degrade the current compliance, as well as making this dependent on the location of the cell in the array. Self-complaint cells (based on a build-in series resistance [33], or by using CRS configuration) may improve that situation.

3 Target specifications for select devices

3.1 General *I-V* requirements.

Current requirements

In its "ON" state, the selector must be able to conduct the read and program currents that are set by the memristive memory element. Typically, program currents are 10x higher than the read current, with minimum read currents of 0.1-1 μA (in order to achieve reasonable read times). This results in program currents of 1-10 μA. For a 10nmx10nm area device, this would result in current densities of 10^6-10^7 A/cm^2. These are huge current densities, difficult to achieve in any type of device, and so this sets a high target for scaled select devices. Further, the select device should be able to conduct these current levels without the need of applying high voltages (> a few Volts) over the device (note that contact resistances may become more important for scaled devices as well [34]).

This current density value is close to the electromigration limit of even metallic conductors and will also result in important resistive drops in the bit and word lines of the memory array. For this reason, programming of memristive memory (sub)arrays has to be done cell by cell (no parallelism).

Voltage requirements

Here, we analyze first the case for a symmetric diode type selector (Type I, Fig. 8), assuming the memory element is in its low resistance state LRS. The selector should be OFF up to a certain voltage V_t applied to the 1S1R cell ($V_t > V_{read}/2$ in $V/2$ scheme, or $V_t > V_{read}/3$ in $V/3$ scheme). As the OFF resistance of the selector is much higher than R_{LRS}, this means that this Vt is also over to the selector and hence it is defining the true selector threshold. At $V=V_{read}$ and $V=V_{write}$, the selector should be ON and V_{read} is shared between the selector and the memory element. The voltage drop over the select device during read and write can be minimized by increasing the *I-V* slope beyond V_t, i.e. by increasing the *I-V* non-linearity. It is of particular interest to minimize the write voltage, in any case so that at $V_{write}/2$ ($V_{write}/3$ in $V/3$ scheme) the voltage over the memory element is smaller than the disturb voltage. Ideally, to minimize the leakage current, $V_{write}/2 < V_t$ ($V_{write}/3 < V_t$ in $V/3$ scheme).

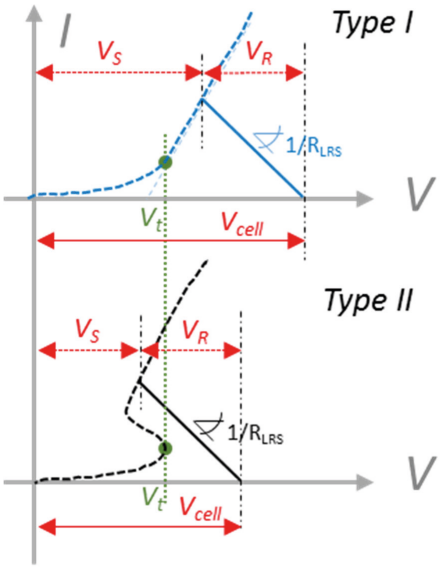


Fig. 8: *1S1R cell voltage distribution during read or write (by load line analysis) over the selector (S) and the memristive element (R), assuming R is in LRS state, for Type I and Type II selector. At constant Vt, smaller cell voltage is required for type II selector due to voltage snapback. V_t is the voltage below which S = OFF (defined by max leakage current).*

Comparison of symmetric diode versus threshold switch

In the case of a threshold switch Type II selector, the situation is different: when one increases the voltage over the threshold voltage, the turn-ON *I-V* characteristic shows negative differential resistance (NDR), and the voltage over the select device "snaps back" to a lower value. As a result, the there is a "snap-forward" of the voltage over the memory element (due to voltage redistribution). If the resultant voltage over the memory element becomes larger than the disturb voltage, erroneously writing of the memory cell may result even during a read operation. This requires a critical balance between threshold voltage and series resistance of the device [35]. On the other hand, due to reduced voltage over the select device in the ON state, total cell voltage during read and write is lowered (Fig. 8, Type II), so this type of selector is better for low-voltage operation.

3.2 Specific requirements based on circuit analysis

More specific quantitative specifications of the selector can be obtained through a detailed circuit analysis of memory array operation (including e.g. line resistances), if the characteristics (and a good circuit model !) for the chosen memristive memory element are available, and given additional specifications as minimum memory window and maximum power dissipation. Different such analysis have been presented, giving more insight in what are the critical select device parameters and what are the trade-off considerations for optimizing the memory performance, see e.g. [36], [37], [38],[39], and [40].

An interesting result is obtained from [35]: based on the given characteristics of a thin HfO_2 based bipolar switching ReRAM element [41], the specifications for both Type I and Type II select devices are determined based on a 1Mbit array at 10nm cell size. (Table I). For a program current of 10 µA, both select devices should have a NL1/2 well above 1000, while the select device type imposes an important difference in cell operation voltage, requiring more than double programming voltage for the Type I selector.

Table I : *Required specifications for Type I and Type II selectors, from 1Mb 1S1R array analysis, given a specific memristive element (RSE) [41]. Program current was set to 10 μA for comparing both selector types, while for Type I selectors the influence of lower current was also evaluated. (From [5], [35])*

Array size:1024x1024		**Type I**		**Type II**
RSE	$I_{switching}$ (μA)	1	10	10
	$V_{set/reset}$ (V)	±1.5	±1.5	±1.5
	$\|V_{disturb}\|$ (V)	0.5	0.5	0.5
	V_{read} (V)	-0.1	-0.1	0.5
Selector	J_{drive} (A/cm^2)	>10^6	>10^7	>10^7
	$\|V_{op}\|$ (V)	>1.5	>2.4	N.A
	NL½	>800	>2000	>5000
	$\|V_{th}\|$ (V)	N.A	N.A	(0.7,1.3)
	R_S (MΩ)	N.A	N.A	0.1$_{(fixed)}$
	I(½V_{op} or ½V_{th}) (nA)	<1.25	<5	<2
1S1R	$\|V_{set/reset}\|$ (V)	>3	>3.9	1.8

Note that so far, only DC analysis have been reported. While matching DC performance is a first requirement, AC performance (including also capacitance analysis) needs to be assessed as well, e.g. to get a more realistic analysis of the power dissipation.

3.3 Operation speed

In order not to compromise the read access time or programming speed of the memory, the select device response to an applied voltage should be faster than that of the memory element itself. With a memory device speed of few ns, this is not an issue for most (especially majority carrier based) electronic devices (bulk MOSFETs and Schottky barrier diodes), but it is a possible concern e.g. for electronic devices relying on generation/recombination of minority carriers (punch-through diodes), and for devices operating through Joule-heating induced insulator to metal phase transitions, where thermal time constants may delay both ON or OFF switching of the device (note that for proper operation, both fast ON and OFF characteristics are required). While different select devices have been proposed, mainly DC characteristics and only in limited cases AC performance data have been reported.

3.4 Reliability and variability

Apart of the standard operation characteristics, also stability of the device characteristics over operation time as well as the effect of select device variability on the cell performance should be checked. It is clear that the endurance of the selector element to read resp. write voltage cycling should at least match the read resp. write endurance of the memory element.

Taking into account the three different variability causes in an 1S1R array (data pattern randomness, selector variability and memristor element variability), it has been found [42] that data pattern randomness is not an important factor as long as the selector limits the leakage currents in the unselected cells, while the selector variability mainly affects the LRS readout current (while, as for the memory element, the smallest read window taking into account the variability in the LRS and HRS distributions directly affects the read margin).

3.5 Scalability

For obtaining high-density (2D) memory arrays, the lateral device dimensions of the select device should scale together with that of the resistive element. As outlined in section **3.1**., this puts increasingly stringent conditions on the operating current density of the selector. Moreover, in a 1S1R stacked cell, the device aspect ratio should be limited, imposing a constraint on the longitudinal device dimensions. This clearly favors thin oxide film tunnel-barrier devices over complex devices as P^+NP^+ punchtrough diodes (see also section **4.1** below). For vertical 3D memory, the total (SSD) device stack "height" is also strongly limited as it directly influences the minimum lateral device pitch.

4 Material, integration, and scaling constraints

4.1 2D, stacked 2D, and 3D memory arrays

While the memristive memory elements considered are typically metal-insulator-metal structures that can be integrated in the back-end process, there are different possible integration schemes for fabricating memory arrays that also depend on front-end or back-end integration of the select device :

(i) 2D memory arrays with the selector integrated in the substrate: this configuration allows for the use of standard Si based transistors or diodes as select devices. While transistors may provide the best possible select devices, they are however typically planar devices that consume Si area resulting in larger memory cells (see section **2.2**). This integration scheme moreover precludes the use of the Si area under the memory array e.g. for integrating the memory periphery. Typically this is an integration scheme well suited for embedded memories where area is less of a concern while the excellent Si device performance aligns with the demand for improved operation.

(ii) Fully back-end integrated 2D and vertically stacked 2D memory arrays: For these configurations, the full cell (memory and select device) is integrated in the back-end. Select devices now can only be made using polycrystalline or amorphous materials, and the temperature budget during device fabrication should be compatible with back-end of line temperatures. Advantages of this integration scheme are the possible incorporation of the peripheral logic circuit underneath the memory arrays and also vertical stacking of multi 2D array levels. However, this multilayer integration scheme is limited to only a few layers, as expensive critical lithography and etch process steps have to be repeated for each memory layer making it no longer an effective cost-scaling process.

(iii) Vertical 3D memory arrays: the limitations of lateral scaling of Flash memories have urged a new vertical integration scheme allowing for cost-effective 3D memory array fabrication,

initially presented by Toshiba as BIt Cost Scaling (BICS) [43]. This concept differs from the vertical stacking of horizontal memory arrays, in that the most critical and expensive process steps are only once applied for all memory layers together. Such vertical 3D memory concepts have been proposed also for memristive memory devices [44]. However, the main issue is that these integration schemes are not compatible with the integration of a separate selector device into each memory cell because (i) lack of space, and (ii) an internal electrode layer would shunt all memory cells of the same vertical column line. While a special vertical chain polycrystalline transistor structure has been proposed [45], a more versatile solution requires self-selecting memory devices (see section **2.3**).

4.2 Integration and scaling issues of Si-based select devices

Material wise, we can divide the select devices in Si based devices (or other classical semiconductor materials as Ge), and those based on other materials (mainly metal-oxides and chalcogenides). A variety of Si based semiconductor devices which have potential as select devices have been developed over time and are well characterized (standard MOSFET as well as bipolar junction transistors, depletion transistors, junction diodes, Schottky barrier diodes, punch-through diodes, etc.), and can be fabricated using standard Si (CMOS) technology. These Si-based select devices can be made in the Si substrate (for 2D arrays only), or can be integrated in the back-end for 2D and also 3D arrays using deposited polycrystalline Si (polysilicon) (see section **4.1** above). Most Si-based select devices, however, critically depend on controlled doping levels and/or profiles. This imposes two important issues:

(i) Even when relative low-temperature deposition of polycrystalline Si is possible, doping atoms require activation at high temperatures that are not compatible with backend integration. (use of Ge may reduce the required temperature, but Ge application is limited due to lack of Ohmic contacts to n-Ge).

(ii) Both internal p-n junctions and Schottky contact barriers require longitudinal device dimensions well above the (unbiased) depletion widths. Further, side depletion effects also limit minimum lateral device dimensions [46].Higher dopant levels decrease these depletion widths but in the limit they may lose their functionality due to tunneling effects and effective Ohmic contact formation.

(iii) At dimensions of a few nm, intermediate doping levels (10^{16}-10^{19} at/cm^3) are impossible to control. Indeed, scaled devices are based either on undoped Si (e.g. for the active region) and on very highly doped regions (e.g. for making the contacts), as e.g. in FinFET transistors.

Both (ii) and (iii) make most semiconductor devices very hard to scale. Further, while these devices are well controlled when fabricated in single-crystal Si (substrate), their performance (as carrier mobility) strongly degrades when formed in polycrystalline Si, which also introduces strong device to device variability. The grain boundaries present also form fast diffusion paths for dopants further limiting dopant control in scaled devices. As a result, use of Si-based select devices is mainly limited to single 2D memory arrays with the selector integrated in the Si substrate. One interesting exception is the use of a thin, undoped, amorphous Si layer in a metal-Si-metal Schottky barrier select device [47].

4.3 Metal-oxide (MO_x) versus classical semiconductor devices

Use of metal oxides facilitates (back-end) integration because of low deposition temperatures. However, it is not obvious to fabricate typical semiconductor devices in metal oxides, because of issues in addition to that mentioned above:

(i) One specific issue is the limitation to form n and p-type regions in a metal oxide by substitutional doping, because in metal oxides dopants can be compensated by charged point defects (metal and oxygen vacancies or interstitials) instead of by charge carriers. For this reason, apart from oxides that cannot be doped (as HfO_2), other metal oxides can only doped to one type (n or p), [48] and for making p-n junctions a heterogenous stack of two different metal oxides has to be used. E.g., typical n-type oxide semiconductors are ZnO, TiO_2, SnO_2, and In_2O_3, while typical p-type oxide semiconductors are limited to NiO and Cu_2O. On the other hand, different from classical (covalently bound) semiconductors, in oxides with strong ionic bonding (as ZnO), doping is effective even in amorphous materials, Also, carrier concentration in metal oxides can be strongly influenced by defects nd e.g., oxygen stoichiometry can be used as an equivalent to doping. The latter effect is crucial for the operation of VCM [25] and interface-controlled [49] memristive devices, but impractical for making junction containing semiconductor devices.

(ii) Secondly, the carrier mobility in metal oxides is typically only a small fraction of that in classical semiconductors, severely limiting the current conduction through the device. The distinguished exception to that is the $In(Ga)ZnO_x$ (IZO and IGZO/GIZO) compounds, forming an n-type semiconductor with a mobility of $>10\ cm^2/V.s$, and this even in the amorphous phase [50].

On the other hand, due to their high bandgap compared to classical semiconductors, metal oxides can be advantageously used in (ultra-thin) Schottky barrier and tunnel diode devices. Note, however, that these properties as high bandgap only hold for the “normal” metal oxides, as the metal “sub-oxides” typically have a much lower bandgap that may be useful for avalanche type of devices, while many sub-oxides also show an insulator to metal phase transition (IMT) [51] that can be used for select device fabrication.

4.4 Other materials for selector devices

(i) Apart from Si and metal oxides, also silicon- and metal-nitrides have been used, e.g.for making Schottky and tunnel diode type select devices.

(ii) Chalcogenide materials as phase-change like material compounds are proposed, mainly for their application as Ovonic threshold switch (OTS) [22]. Because of material and process compatibility, this material may in particular be suited for a select device for phase-change memory.

5 Overview of proposed select devices

Section **2.2** introduced the main types of select devices, based on their general *I-V* characteristic. In this section, for each of these categories of select devices, an overview of the different proposed selectors is given, classified according to their physical operation mechanism and the type of material (see also section **4**) used.

5.1 Select Transistors

Table II gives an overview of proposed transistor-type selectors. Note that the polycrystalline Si transistor as proposed by [16] and [45] is restricted to a specific cell configuration (chain cells) and memristor type (phase-change cell); while the GIZO transistor is proposed as common WL or BL selector in combination with (oxide) diodes for the individual cell selection [17].

Table II : Transistor select devices

Basic material	*Device type*	*Material details*	*References*
Si	Planar MOSFET	1-Xtal Si substrate	[52]
	Vertical gate-all-around MOSFET	Si nanopillar	[53]
	Vertical bipolar junction transistor (BJT)	1-Xtal Si substrate	[15], [54]
	Chain-type MOSFET	Polycrystalline Si	[16], [45]
Metal Oxide	Thin film depletion-mode FET	Amorphous Ga_2O_3-In_2O_3-ZnO (GIZO)	[17]

5.2 Diodes

Reported diode selectors are summarized in Table III. Apart from single-crystal and polycrystalline Si-based p-n diodes, metal oxide based junction and Schottky-barrier diodes are reported (Fig. 9). High current density and excellent rectification properties of the low-temperature amorphous GIZO diode, however for a 100nm thick film [55].

Table III: Vertical Diodes

Basic material	*Device type*	*Material details*	*References*
Si	p-n junction diode	Epitaxial Si	[58]
		Polycrystalline Si	[18]
Metal Oxide	p-n junction diode	p-NiO/n-TiO_2	[19] [56]
		p-CuO_x/n-$InZnO_x$	[59]
	Schottky diode	Ti/TiO_2/Pt	[57] [60]
		Ag/ZnO/Ti/Au	[61]
		a-IGZO	[55]

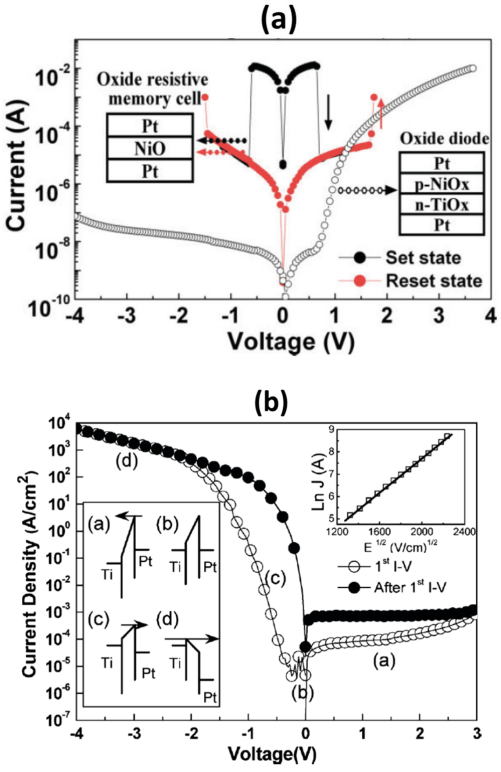


Fig. 9: *(Unipolar) metal-oxide diodes :*
(a) p-NiO_x/n-TiO_x junction diode [56];
(b) . Ti/TiO_2/Pt Schottky-barrier diode [57]

5.3 Bipolar two-terminal selectors:

Type I : Symmetric Diodes

Symmetric diode selector characteristics (Table IV) can be obtained using semiconductor devices as back-to-back Schottky barrier diodes and N^+PN^+ (P^+NP^+) punch-through diodes, Fig. 10.

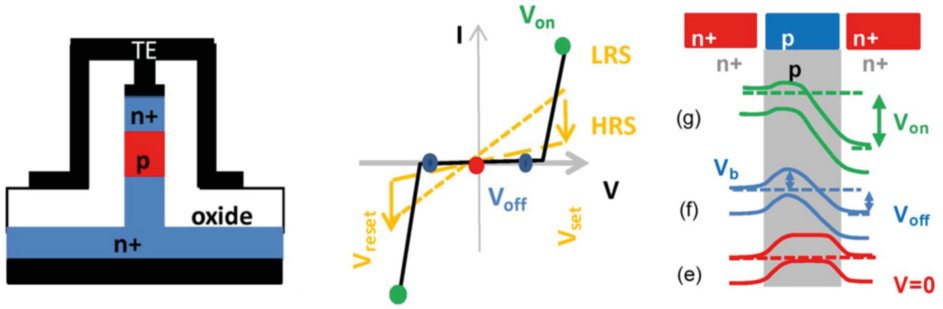


Fig. 10: *Device structure and operation schematic of symmetric vertical N^+PN^+ punch-trough diode [62]*

Much more interesting for scaled devices, however, are symmetric diodes based on tunneling through thin insulator films, due to the intrinsic strong non-linear *I-V* of the tunneling process. This non-linearity can be further increased by using engineered barriers (crested barrier [63] and Variot [64] devices), that are obtained by stacking of different insulator layers (Fig. 11). A special case is the MIEC device [65], based on Cu motion in a (not-disclosed) mixed ionic electronic conduction material. (Fig. 12).

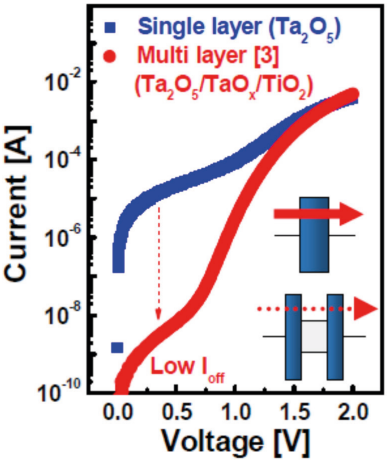


Fig. 11: *Effect of barrier engineering using multi-layer tunnel device [66]*

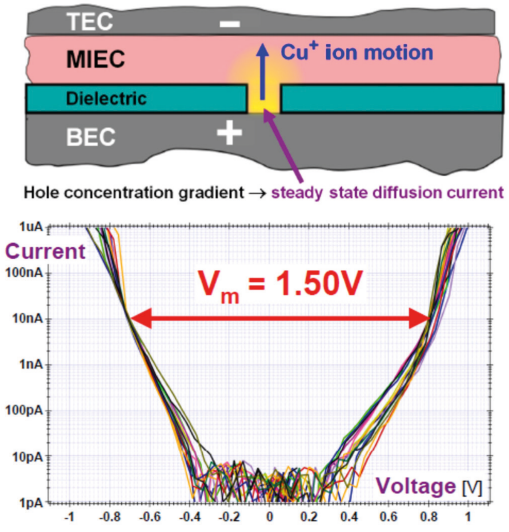


Fig. 12: *Operation mode and characteristic of MIEC device [65], [67]*

An overview of proposed Symmetric diode structures can be found in Table IV:

Table IV: Symmetric bipolar diodes

Basic material	*Device type*	*Material details*	*References*
Si	Punch-through NPN diodes	Epitaxial Si	[62]
		Polycrystalline Si	[68]
	Back–to-back Schottky Barrier diodes	Ta/P-type poly-Si/Ta	[69]
	Single layer MSM tunneling diode	Undoped thin amorphous Si	[47]
SiNx	Single layer tunneling diode (M/SiNx/M**)**	TaN/SiN_x/TaN	[70]
Si/SiNx	Tunnel device with engineered barrier: Thin Si Injector (TSI)	a-S/SiN/a-Si	[71]
Metal Oxide	Punch through NPN diodes	CoO_x/IGZO/CoO_x	[72]
	Single layer MIM Tunneling Diodes	TiN/Ta_2O_5/TiN	[20]
		Ni/TiO_2/Ni	[73], [74]
		Pt/TiO_2/TiN	[75]
		TaN/SiN_x/TaN	[70]
	Tunnel devices with engineered barriers	TaO_x/TiO_2/TaO_x (Varistor) :	[76]
		Pt/Ta_2O_5/TaO_x/TiO_2/Pt	[77]
MIEC	Mixed Electron-Ion Conductor (MIEC) device	undisclosed	[65]

Type II: Threshold Switches

The main types of threshold switching devices are:

(i) Latch-up N^+PN^+ diode : different from a punch-through diode, this device shows NDR and threshold switching characteristics, see Fig. 13.

(ii) Insulator to Metal phase Transition (IMT) (this has also been called Mott transition), occurring in a number of metal sub-oxides as VO_2 and NbO_2.[23],[51]. Important here is that the transition temperature T_c should be much higher than the maximum operation temperature, which is a problem for VO_2 (Tc = 340 K = 67 ^oC only), but OK for NbO_2 (Tc=1070 K, [51]).

(iii) Ovonic Threshold Switching device (OTS) [22]: based on electronic switching in amorphous chalcogenide materials. These materials are similar to Phase Change Memory (PCM) materials, albeit with a higher crystallization temperature T_c. Indeed, PCM materials also show threshold switching, but cannot avoid Joule heating induced crystallization due to their relatively low T_c. Different theories have been proposed to explain this threshold switching phenomenon : thermal [78], Impact Ionization and SHR recombination [79]

[80, 81], trap-limited conduction [82] [83]; and even the formation of instable crystalline nuclei.

(iv) Electronic filamentation (non-IMT related threshold switching) in metal oxides: by limiting the current compliance during forming, a volatile switching has been observed in sub-oxides as TaO_x and TiO_x. [84] [85] It is argued that the forming process is a two-step process, starting with (reversible) electronic filamentation, and only above a certain current (/Joule heating) permanent damage (i.e. oxygen vacancy creation) causes an irreversible conducting filament to form, see Fig. 14. (in principle this process is very similar to the electronic switching in OTS devices). These devices have been proposed for making voltage controlled oscillators enabling compact synthetic neurons, but they could be used as selector device as well.

(v) Avalanche breakdown in small bandgap Mott materials : in "true"Mott materials as $V_{2-x}Cr_xO_3$, a volatile resistive switching is observed, see Fig. 15. Correlation of the threshold voltage to the bandgap suggests that this volatile switching is induced by an avalanche process [86],[87]. Again, this is a (filamentary) electronic switching process, similar to (iii), and (iv).

(vi) Unstable (ionic defect) filament: both in ECM and VCM cells, volatile switching has been observed that has been related to unstable cation resp. oxygen vacancy defects. Different mechanisms have been proposed (mechanical stress in ECM cells, change in energy band structure in Ni deficient NiO, etc.). An (extreme) fingerprint of this kind of threshold switching is the strong hysteresis of the switching, with an ON state of the selector that is stable down to very small biases (~0 V), see Fig. 16. This is also the suggested operation mechanism of the Vacuum Threshold Switch device [88].

(vii) Field assisted superlinear threshold switch (FAST): this device shows a clear threshold switch and has promising selector characteristics, but the material and the mechanism is not disclosed [89]

Table V lists the different proposed threshold switching devices.

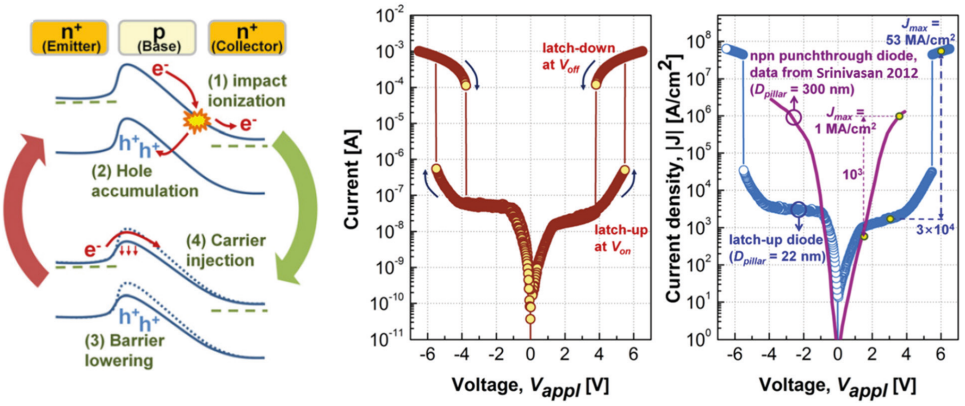


Fig. 13: *Mechanism and threshold switching curves of a latch-up N^+PN^+ diode, compared with the N^+PN^+ punch-through diode of Fig. 10. [90]*

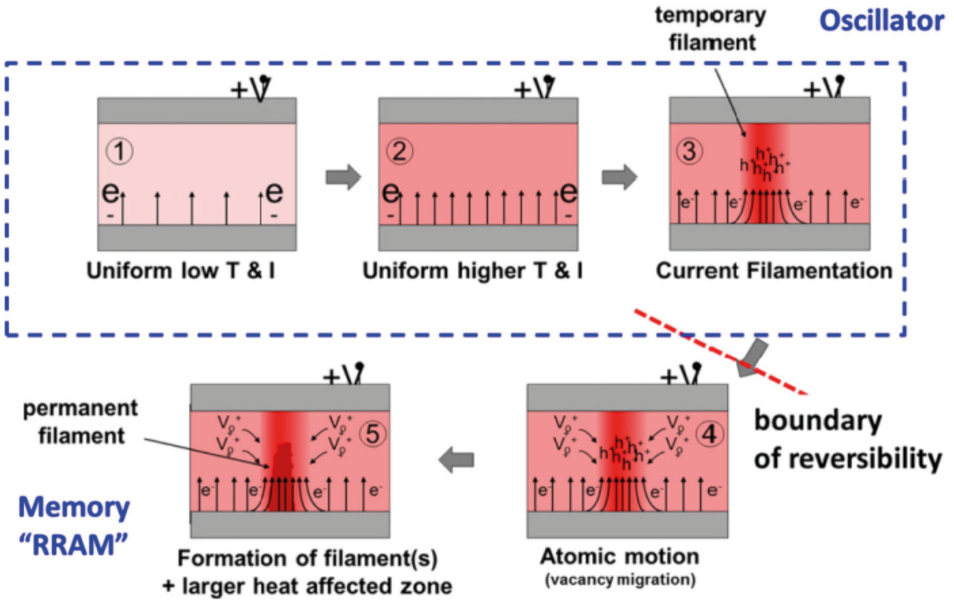


Fig. 14: *Schematic of .boundary between reversible electronic filament formation and irreversible defect filament formation[85].*

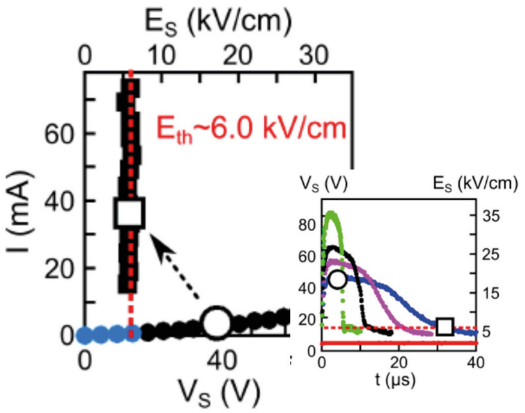


Fig. 15: *Pulse-induced volatile resistive switching in $V_{2-x}Cr_xO_3$. Mott oxide. .The constant holding field corresponds to the avalanche threshold in the material.correlated points correspond to same pulse amplitude [91]*

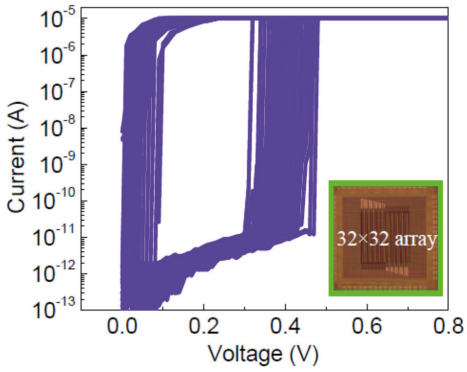


Fig. 16: *Unstable filament shows volatile switching characteristics with large hysteresis, with an ON state of the selector that is stable down to very small biases (~0 V), [92] - compare e.g. with the TS curve of Fig. 13 ...*

Table V: Threshold Switching bipolar diodes

Basic material	*Device type*	*Material details*	*References*
Si	Latch-up NPN diode	1 Xtal Si pillar	[90]
Chalcogenides	Ovonic Threshold Switch (OTS)	undisclosed	[93]
		$GeTe_6$	[94]
		AsTeGeSiN	[21]
		doped chalcogenide	[95]
		CuGeS	[96]
Metal Oxides	Insulator to Metal Transitions (IMT) (~ Mott)	VO_2	[97]
		NbO_2	[98]
	Electronic filamentation	TiO_x, TaO_x	[84] [85]
	Avalanche induced volatile switching in low bandgap true Mott materials	$V_{2-x}Cr_xO_3$	[86]
	Unstable filament	NiO (VCM)	[99]. [100]
		Threshold Vacuum Switch (TVS), WO_x (VCM)	[88]
		Cu/TaO_x/Pt (ECM)	[101]
		Cu-doped HfO_2 (ECM)	[92]
Not disclosed	Field assisted superlinear threshold switch (FAST)	undisclosed	[89]

Comparison of bipolar selector characteristics

Fig. 17 and Table VI compare the main Type I and Type II bipolar selector devices [5], [35]. MIEC and Varistor selectors have the best performance in terms of high non-linearity, however, their operating voltage range is too low for many (esp. VCM) memristive devices. Among the Type II selectors, FAST shows best nonlinear characteristics and proper threshold voltage, but the low ON resistance may lead to potential read disturb.

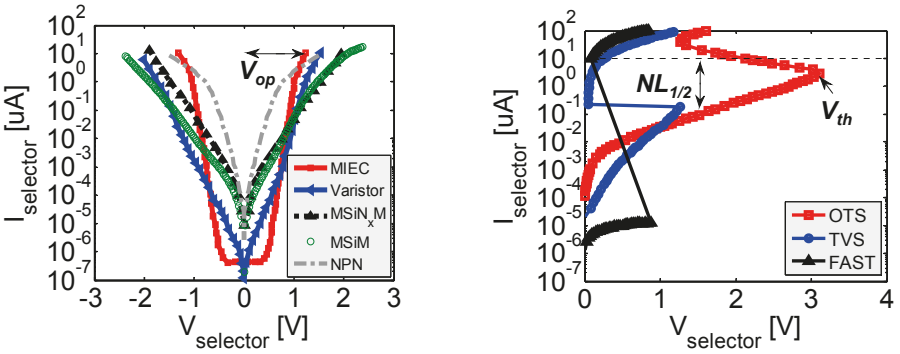


Fig. 17: *Experimental I-V characteristics of (a) type-I selectors and (b) type-II selectors. For all the selectors, we estimate the current assuming the selector is able to deliver 10μA drive current and a constant current-density [5]. Considered type-I selectors are MSM [47], Varistor [76], M/SiNx/M [70], silicon NPN [62]and MIEC [65]; and type-II selectors :OTS [21], FAST [89],and TVS [88](see Tables IV and V).*

Table VI : Comparison of bipolar selectors

Type I	***V_{op} (V)***	***NL½***	***Slope(mV/dec)***	***J_{drive}(A/cm^2)***
MIEC	1.2	8*10^4	85	10^7
Varistor	1.6	10^4	282	10^7
MSM	2.4	260	330	10^6
MSiN$_x$M	2	150	330	10^5
NPN	1.6	100	219	10^6
Type II	***V_{th} (V)***	***NL½***	***R_S (kΩ)***	***J_{drive}(A/cm^2)***
FAST	0.8	10^7	8	>5*10^6
TVS	1.3	2360	11	>10^7
OTS	3.1	110	9	>10^7

5.4 Self-Selecting Devices

Finally, Table VII gives an overview of different proposed SSD's. Main concepts include:

(i) Hybrid devices : combining bipolar switching stoichiometric metal oxide layer with an IMT sub-oxide layer (Fig. 18) or, combining a metal oxide switching layer with an OTS layer; in both cases without an intermediate electrode.

(ii) Non-linear memristive devices: by adding a (fixed) non-linear tunneling layer to the switching layer.

(iii) Devices with a varying-width tunneling (or Schottky) barrier (Fig. 19).

(iv) Complementary switching (CS) devices (we consider here only single devices, without intermediate electrode, i.e. no CRS devices, see section **3.2**).

Note that switching in type (ii) and (iii) devices is different from the usual filamentary switching (as in type (i) and (iv) devices), and may be areal switching.

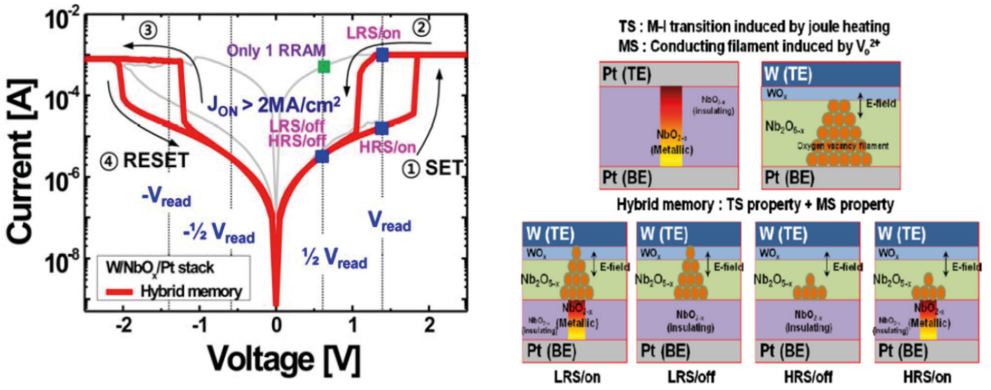


Fig. 18: *Hybrid VCM/IMT SSD : I-V characteristics and device operation schematic [98].*

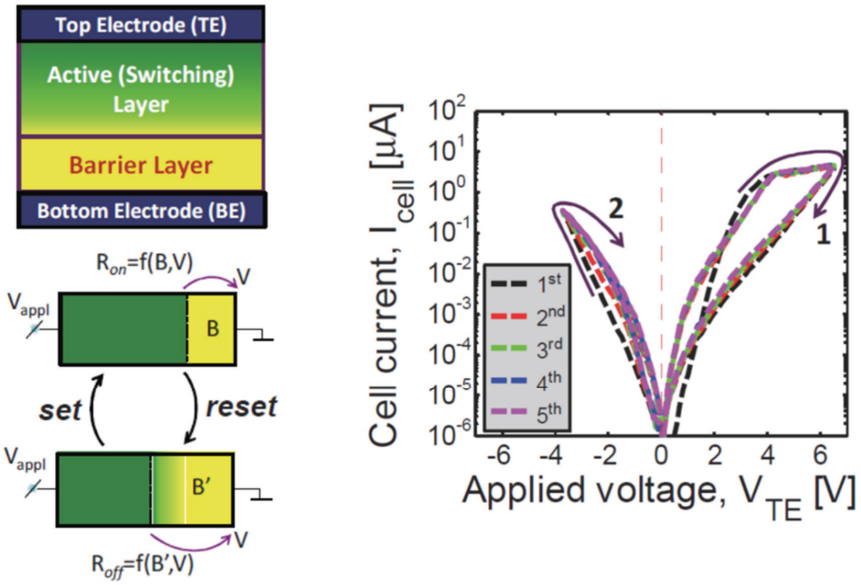


Fig. 19: *a-VMCO (Vacancy-Modulated Conductive Oxide) cell operation based on varying tunnel barrier thickness and switching characteristic showing non-linear ON state. [102]*

Table VII: Self-selecting devices

Device Type		*Material details*	*References*
Hybrid devices	combining VCM with IMT layer	Nb_2O_5/NbO_2	[98] [103]
	Combining VCM with OTS layer	$CuGeS/HfO_2$	[96]
Non-linear devices	Adding fixed tunnel barrier to VCM layer	2 TMO's with barriers (non disclosed)	[104]
		TaO_x/TiO_{2-x}	[105]
		TiO_x/TiO_y bilayer	[106]
	Charge trapping effect	Ta_2O_5/TiO_x	[107]
Variable tunneling barrier	Vacancy-Modulated Conduction Oxide ReRAM (VMCO)	$TiN/Al_2O_3/TiO_2/TiN$	[108]
	a-VMCO	$a\text{-}Si/TiO_2$	[102]
	Conductive Metal Oxide $(CMOX)^{TM}$ cell (O defects moving in tunnel barrier and changing trapped charge)	Pt/PCMO/ tunnel oxide/Pt	[109], [110]
Variable Schottky barrier	Formation/reduction of interfacial AlO_x layer	Pt/PCMO/Al/Pt	[111]
Complementary Switching	(Single cells)	$W/Nb_2O_{5-x}/NbO_y/Pt$	[112]
		TaO_x	[113]
			[114]
		$Pt/TiO_{2-x}/TiN_xO_y/TiN$	[115]

6 Conclusions

In order to fabricate functional high density crossbar arrays of memristive devices (for applications beyond embedded memories using 1T1R arrays), it is necessary to incorporate a two-terminal selector element in each cell (or, alternatively, to build this selector functionality into the memristive device itself). Circuit simulation is hereby an indispensable tool to determine the required selector specifications for a given memristor device.The obtained requirements are quite hard to fulfil: e.g., for a bipolar switching memristive device, a huge current density

(10 A/cm^2) has to be combined with a high nonlinearity (> 1000), while being compatible with program and read voltages of the memristive device (and overall low voltage operation of the cell). Further, switching speed and reliability (cyclability) and variability requirements have to be met, too, while material and process compatibility, as well as scalability impose further constraints on the selector device and material

While main research focus during the last ten years was on the memristive device itself, these demands have spurred strong selector research during the last couple of years. As an important result, quite a large number of such select devices have been proposed in recent literature, some of them enabling first dense 1S1R demonstrator arrays. As the R&D of the select device has catched up with that of the memristive device, we may soon see some real commercial applications launched.

Acknowledgments

I am particularly indebted to Dr. Leqi Zhang, my former PhD student at imec/KULeuven (Leuven, Belgium) for his excellent research work on the selector element for resistive memory [5] that constituted a useful guideline and inspiration for writing.this chapter.

References

[1] G. Kar *et al.*, "Process-improved RRAM cell performance and reliability and paving the way for manufacturability and scalability for high density memory application," *2012 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. 157-158, 2012.

[2] E.P.G. Wright, US patent 2,667,542, Jan. 26, 1954.

[3] A. Flocke, T.G. Noll, C. Kugeler, C. Nauenheim, and R. Waser, "A fundamental analysis of nano-crossbars with non-linear switching materials and its impact on TiO_2 as a resistive layer," *Proceedings of the 8th IEEE Conference on Nanotechnology*, pp. 319-322, 2008.

[4] A. Flocke and T. G. Noll, "Fundamental analysis of resistive nano-crossbars for the use in hybrid Nano/CMOS-memory," *Proceedings of the 33rd European Solid-State Circuits Conference*, pp. 328-331, 2007.

[5] L. Zhang, "Study of the Selector Element for Resistive Memory", Ph.D.thesis, KULeuven, 2015.

[6] H. Schroeder, V. V. Zhirnov, R. K. Cavin, and R. Waser, "Voltage-time dilemma of pure electronic mechanisms in resistive switching memory cells," *J. Appl. Phys.*, vol. 107, pp. 054517/1-8, 2010.

[7] E. Linn, R. Rosezin, C. Kügeler, and R. Waser, "Complementary Resistive Switches for Passive Nanocrossbar Memories," *Nat. Mater.*, vol. 9, pp. 403-406, 2010.

[8] J. Liang and H.-S. P. Wong, "Size limitation of cross-point memory array and its dependence on data storage pattern and device parameters," *2010 IEEE International Interconnect Technology Conference (IITC)*, 2010.

[9] A. Chen, “Accessibility of nano-crossbar arrays of resistive switching devices,” *11th IEEE International Conference on Nanotechnology* (*IEEE-Nano)*, p. 1767, 2011.

[10] L. Zhang *et al.*,“On the Optimal ON/OFF Resistance Ratio for Resistive Switching Element in One-Selector One-Resistor Crosspoint Arrays,” *IEEE Electron Device Lett.*, vol. 36, pp. 570-572, 2015.

[11] R. Waser, V. Rana, S. Menzel, and E. Linn, “Energy-efficient Redox-based Non-Volatile Memory Devices and Logic Circuits,” *3rd Berkeley Symposium on Energy Efficient Electronic Systems*, pp. 1-2, 2013.

[12] G. Burr *et al.*, “Access devices for 3D crosspoint memory,” *J. Vac. Sci. Technol. B*, vol. 32, pp. 040802, 2014.

[13] A. Chen, “A Comprehensive Crossbar Array Model With Solutions for Line Resistance and Nonlinear Device Characteristics,” *IEEE Trans. Electron Devices*, vol. 60, pp. 1318 - 1326, 2013.

[14] C. Xu, X. Dong, N.OP Jouppi, and Y. Xie, “Design implications of memristor-based RRAM cross-point structures,” *2012 Design, Automation & Test in Europe (DATE)*, pp., 2011, 2012.

[15] C.-H. Wang *et al.*, “Three-Dimensional 4F2 ReRAM Cell with CMOS Logic Compatible Process,” *2010 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 4, 2010.

[16] Y. Sasago *et al.*,“Phase-change memory driven by poly-Si MOS transistor with low cost and high-programming gigabyte-per-second throughput,” *2011 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. 96 - 97, 2011.

[17] M.J. Lee *et al.*, “Stack Friendly All-Oxide 3D RRAM using GaInZnO Peripheral TFT realized over Glass Substrates,” *2008 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 1-4, 2008.

[18] Y. Sasago *et al.*, “Cross-point phase change memory with 4F(2) cell size driven by low-contact-resistivity poly-Si diode,” *2009 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. 24-25, 2009.

[19] I. G. Baek *et al.*, “Multi-layer cross-point binary oxide resistive memory (OxRRAM) for post-NAND storage application,” *2005 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 769-772, 2005.

[20] B. Govoreanu *et al.*, “High-Performance Metal-Insulator-Metal Tunnel Diode Selectors,” *IEEE Electron Device Lett.*, vol. 35, pp. 63-65, 2014.

[21] S. Kim *et al.*, “Performance of threshold switching in chalcogenide glass for 3D stackable selector,” *2013 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. T240 - T241, 2013.

[22] S. R. Ovshinsky, “Reversible electrical switching phenomena in disordered structures,” *Phys. Rev. Lett.*, vol. 21, pp. 1450-3, 1968.

[23] M. Imada, A. Fujimori, and Y. Tokura, “Metal-insulator transitions,” *Rev. Mod. Phys.*, vol. 70, pp. 1039-1263, 1998.

[24] S. Tappertzhofen *et al.*, "Capacity based Nondestructive Readout for Complementary Resistive Switches”, *Nanotechnology*, vol. 22, pp.395203/1-7, 2011.

[25] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-Based Resistive Switching Memories - Nanoionic Mechanisms, Prospects, and Challenges," *Adv. Mater.*, vol. 21, pp. 2632-2663, 2009.

[26] R. Rosezin *et al.*, "Integrated Complementary Resistive Switches for Passive High-Density Nanocrossbar Arrays," *IEEE Electron Device Lett.*, vol. 32, pp. 191-193, 2011.

[27] J. Lee *et al.*, "Diode-less nano-scale ZrO_x/HfO_x RRAM device with excellent switching uniformity and reliability for high-density cross-point memory applications," *2010 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 451-453, 2010.

[28] D. J. Wouters *et al.*, "Analysis of Complementary RRAM Switching," *IEEE Electron Device Lett.*, vol. 33, pp. 1186 - 1188, 2012.

[29] T. Breuer *et al.*, "Low-current operations in 4F^2 -compatible Ta2O5 -based complementary resistive switches," *Nanotechnology*, vol. 26, pp. 415202, 2015.

[30] F. Nardi, C. Cagli, S. Spiga, and D. Ielmini, "Reset Instability in Pulsed-Operated Unipolar Resistive-Switching Random Access Memory Devices," *IEEE Electron Device Lett.*, vol. 32, pp. 719-721, 2011.

[31] D. Ielmini, F. Nardi, and C. Cagli, "Universal Reset Characteristics of Unipolar and Bipolar Metal-Oxide RRAM," *IEEE Transactions on Electron Devices*, vol. 58, pp. 1-8, 2011.

[32] K. Kinoshita *et al.*, "Reduction of reset current in NiO-ReRAM brought about by ideal current limiter," *2007 22nd IEEE Non-Volatile Semiconductor Memory Workshop*, pp. 66-7, 2007.

[33] Y-S. Fan *et al.*, "Direct Evidence of the Overshoot Suppression in Ta2O5-Based Resistive Switching Memory with an Integrated Access Resistor," *IEEE Electron Device Lett.*, vol. 36, pp. 1027-1029, 2015.

[34] K. Martens *et al.*, "Record low contact resistivity to n-type Ge for CMOS and memory applications," *2010 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 18.4.1 - 18.4.4, 2010.

[35] L. Zhang *et al.*, "One-Selector One-Resistor Cross-Point Array With Threshold Switching Selector," *IEEE Trans. Electron Devices*, vol. 62, pp. 3250-3257, 2015.

[36] Y. Deng *et al*., "RRAM Crossbar Array With Cell Selection Device: A Device and Circuit Interaction Study," *IEEE Trans. Electron Devices*, vol. 60, pp. 719-726, 2013.

[37] A. Chen, "Comprehensive methodology for the design and assessment of crossbar memory array with nonlinear and asymmetric selector devices," *2013 IEEE International Electron Devices Meeting (IEDM) Technical*, pp. 30.3.1-30.3.4., 2013.

[38] S. Kim, J. Zhou, and W.D. Lu, "Crossbar RRAM Arrays: Selector Device Requirements During Write Operation," *IEEE Trans. Electron Devices*, vol. 61, pp. 2820-2826, 2014.

[39] J. Zhou, K-H. Kim, and W. Lu, "Crossbar RRAM Arrays: Selector Device Requirements During Read Operation," *IEEE Trans. Electron Devices*, vol. 61, pp. 1369-1376, 2014.

[40] L. Zhang *et al.*, "Selector design considerations and requirements for 1S1R RRAM crossbar array," 2014 *6th IEEE International Memory Workshop (IMW)*, pp. 4. 2014.

[41] B. Govoreanu *et al.*, "Performance and reliability of Ultra-Thin HfO2-based RRAM (UTO-RRAM)," *2013 5th IEEE International Memory Workshop (IMW)*, pp. 48-51, 2013.

[42] L. Zhang *et al.*, "Cell Variability Impact on the One-Selector One-Resistor Cross-Point Array Performance," *IEEE Trans. Electron Devices*, vol. 62, pp. 3490-3497, 2015.

[43] H. Tanaka *et al.*, "Bit cost scalable technology with punch and plug process for ultra high density Flash memory," *2007 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers,* pp. 14–15, 2007.

[44] I. Baek *et al.*, "Realization of vertical resistive memory (VRRAM) using cost effective 3D process," *2011 IEEE International Electron Devices Meeting (IEDM) Technical Digest Technical Digest*, pp. 31.8.1-31.8.4., 2011.

[45] M. Kinoshita *et al.*, "Scalable 3-D vertical chain-cell-type phase-change memory with 4F2 poly-Si diodes," *2012 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers),pp.35-36*, 2012.

[46] V. V. Zhirnov, R. Meade, R. K. Cavin, and G. Sandhu, "Scaling limits of resistive memories," *Nanotechnology*, vol. 22, pp. 254027/1-21, 2011.

[47] L. Zhang *et al.*, "Ultrathin Metal/Amorphous-Silicon/Metal Diode for Bipolar RRAM Selector Applications," *IEEE Electron Device Lett.*, vol. 35, pp. 199-201, 2014.

[48] J. Robertson and S.J. Clark, "Limits to doping in oxides," *Phys. Rev. B: Condens. Matter*, vol. 83, pp. 075205 1-7, 2011.

[49] A. Sawa, "Resistive switching in transition metal oxides," *Mater. Today*, vol. 11, pp. 28-36, 2008.

[50] H. Yabuta *et al.*, "High-mobility thin-film transistor with amorphous InGaZnO4 channel fabricated by room temperature rf-magnetron sputtering," *Appl. Phys. Lett.*, vol. 89, pp. 112123/1-3, 2006.

[51] F. A. Chudnovskii, L. L. Odynets, A. L. Pergament, and G. B. Stefanovich, "Electroforming and switching in oxides of transition metals: the role of metal-insulator transition in the switching mechanism," *J. Solid State Chem.*, vol. 122, pp. 95-9, 1996.

[52] M. Terai *et al.*, "High thermal robust ReRAM with a new method for suppressing read disturb," *2011 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. 50, 2011.

[53] X. P. Wang *et al.*, "Highly compact 1T-1R architecture (4F 2 footprint) involving fully CMOS compatible vertical GAA nano-pillar transistors and oxide-based RRAM cells exhibiting excellent NVM properties and ultra-low power operation," *2012 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, p. 20.6 (4 pp.), 2012.

[54] M-F. Chang *et al.,* "Area-Efficient Embedded Resistive RAM (ReRAM) Macros Using Logic-Process Vertical-Parasitic-BJT (VPBJT) Switches and Read-Disturb-Free Temperature-Aware Current-Mode Read Scheme," *IEEE J. Solid-State Circuits*, vol. 49, pp. 908-916, 2014.

[55] A. Chasin *et al.*, "High-Performance a-IGZO Thin Film Diode as Selector for Cross-Point Memory Application," *IEEE Electron Device Lett.*, vol. 35, pp. 642-644, 2014.

[56] M.-J. Lee *et al.*, "A Low-Temperature-Grown Oxide Diode as a New Switch Element for High-Density, Nonvolatile Memories**," *Adv. Mater.*, pp. 73, 2007.

[57] Jiun-Jia Huang, Chih-Wei Kuo, Wei-Chen Chang, and Tuo-Hung Hou, "Transition of stable rectification to resistive-switching in Ti/TiO[sub 2]/Pt oxide diode," *Appl. Phys. Lett.*, vol. 96, pp. 262901, 2010.

[58] J. H. Oh *et al.*, "Full integration of highly manufacturable 512Mb PRAM based on 90nm technology," *2006 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 515-518, 2006.

[59] B. S. Kang *et al.*, "High-current-density CuOx/InZnOx thin-film diodes for cross-point memory applications," *Adv. Mater.*, vol. 20, pp. 3066-3069, 2008.

[60] W. Y. Park *et al.*, "A Pt/TiO_2/Ti Schottky-type selection diode for alleviating the sneak current in resistance switching memory arrays," *Nanotechnology*, vol. 21, pp. 195201/1-4, 2010.

[61] G. Tallarida *et al.*, "Low temperature rectifying junctions for crossbar non-volatile memory devices," *2009 IEEE International Memory Workshop* (IMW), pp. 6-8, 2009.

[62] V. S. S. Srinivasan *et al.*, "Punchthrough-Diode-Based Bipolar RRAM Selector by Si Epitaxy," *IEEE Electron Device Lett.*, vol. 33, pp. 1396-1398, 2012.

[63] K. K. Likharev, "Layered tunnel barriers for nonvolatile memory devices," *Appl. Phys. Lett.*, vol. 73, pp. 2137-9, 1998.

[64] B. Govoreanu, P. Blomme, M. Rosmeulen, J. Van-Houdt, and K. De-Meyer, "VARIOT: a novel multilayer tunnel barrier concept for low-voltage nonvolatile memory devices," *IEEE Electron Device Letters, USA*, vol. 24, pp. 99-101, 2003.

[65] K. Gopalakrishnan *et al.*, "Highly-Scalable Novel Access Device based on Mixed Ionic Electronic Conduction (MIEC) Materials for High Density Phase Change Memory (PCM) Arrays," *2010 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. 205-206, 2010,

[66] J. Woo *et al.*, "Electrical and Reliability Characteristics of a Scaled (~30nm) Tunnel Barrier Selector (W/Ta2O5/TaOx/TiO2/TiN) with Excellent Performance (JMAX > 107A/cm2)," *2014 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers,* pp. 1 - 2, 2014.

[67] G. W. Burr *et al.*, "Large-scale (512kbit) integration of Multilayer-ready Access-Devices on Mixed-Ionic-Electronics-Conduction(MIEC) at 100% yield," *2012 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. 41-42, 2012.

[68] M. H. Lee *et al*., "Reliability of Ambipolar Switching Poly-Si Diodes for Cross-Point Memory Applications," *2011 Device Research Conference (DRC)*, pp. 89-90, 2011.

[69] Y-S. Park, G-H. Kil, and Y-H. Song, "Bidirectional Two-Terminal Switching Device Using Schottky Barrier for Spin-Transfer-Torque Magnetic Random Access Memory," *Jpn. J. Appl. Phys.*, vol. 51, pp. 106501-1 to 5, 2012.

[70] A. Kawahara *et al*., "An 8 Mb Multi-Layered Cross-Point ReRAM Macro With 443 MB/s Write Throughput," *Solid-State Circuits, IEEE Journal of*, pp. 178-185, 2013.

[71] B. Govoreanu *et al.*, "Thin-Silicon Injector (TSI): An All-Silicon Engineered Barrier, Highly Nonlinear Selector for High Density Resistive RAM Applications," *IMW 2015*, pp., 2015.

[72] Y.C. Bae *et al.*, "All oxide semiconductor-based bidirectional vertical p-n-p selectors for 3D stackable crossbararray crossbararray," *Scientific Reports*, pp. 5:13362, 2015.

[73] J.-J. Huang *et al.*, "One Selector-One Resistor (1S1R) Crossbar Array for High-density Flexible Memory Applications," *2011 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp.734-736, 2011.

[74] J.-J. Huang, Y.-M. Tseng, C.-W. Hsu, and T.-H. Hou, "Bipolar Nonlinear Ni/TiO2/Ni Selector for 1S1R Crossbar Array Applications," *IEEE Electron Device Lett.*,pp. 1-3, 2011.

[75] J. Shin *et al.*, "TiO(2)-based metal-insulator-metal selection device for bipolar resistive random access memory cross-point application," *J. Appl. Phys.*, vol. 109, pp. 033712/1-4, 2011.

[76] W. Lee *et al.*, "High Current Density and Nonlinearity Combination of Selection Device Based on $TaO_x/TiO_2/TaO_x$ Structure for One Selector–One Resistor Arrays," *ACS Nano*, vol. 6, pp. 8166-8172, 2012.

[77] J. Woo *et al*., "Multi-layer tunnel barrier (Ta2O5/TaO x/TiO2) engineering for bipolar RRAM selector applications," *2013 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. T168-T169, 2013.

[78] D. M. Kroll, "Theory of Electrical Instabilities of Mixed Electronic and Thermal Origin," *Physical Review B*, vol. 9, pp. 1669-1706, 1974.

[79] D. Adler, H. K. Henisch, and N. Mott, "The mechanism of threshold switching in amorphous alloys," *Rev. Mod. Phys.*, vol. 50, pp. 209-20, 1978.

[80] D. Adler, M. S. Shur, M. Silver, and S. R. Ovshinsky, "Threshold switching in chalcogenide-glass thin films," *J. Appl. Phys.*, vol. 51, pp. 3289-309, 1980.

[81] A. Pirovano, A. L. Lacaita, A. Benvenuti, F. Pellizzer, and R. Bez, "Electronic switching in phase-change memories," *IEEE Transactions on Electron Devices, USA*, vol. 51, pp. 452-9, 2004.

[82] D. Ielmini and Y. Zhang, "Analytical model for subthreshold conduction and threshold switching in chalcogenide-based memory devices," *J. Appl. Phys.*, vol. 102, pp. 054517-1-3, 2007.

[83] D. Ielmini, "Threshold switching mechanism by high-field energy gain in the hopping transport of chalcogenide glasses," *Phys. Rev. B*, vol. 78, pp. 035308, 2008.

[84] A.A.Sharma *et al.*, "HIgh Performance, Integrated 1T1R Oxide-based Oscillator: Stack Engineering for Low-Power Operation in Neural Networkl Applications," *2015 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, p. T186, 2015.

[85] A.A:Sharma, J.A.Bain, and J.A.Weldon, "Phase Coupling and Control of Oxide-Based Oscillators for Neuromorphic Computing," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 1, pp. 5866, 2015.

[86] P. Stoliar *et al.*, "Nonthermal and purely electronic resistive switching in a Mott memory," *Phys. Rev. B: Condens. Matter*, vol. 90, pp. 45146/1-, 2014.

[87] V. Guiot, E. Janod, B. Corraze, and L. Cario, "Control of the Electronic Properties and Resistive Switching in the New Series of Mott Insulators GaTa4Se8-yTey ($0 <= y <= 6.5$)," *Chem. Mater*., vol. 23, pp. 2611-2618, 2011.

[88] C.H. Ho *et al.*, "Threshold Vacuum Switch (TVS) on 3D-Stackable and 4F2 Cross-Point Bipolar and Uni-polar Resistive Random Access Memory," *2012 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 2.8.1 - 2.8.4, 2012.

[89] S.H. Jo, T. Kumar, S. Narayanan, W.D. Lu, and H. Nazarian, "3D-stackable Crossbar Resistive Memory based on Field Assisted Superlinear Threshold (FAST)," *2014 IEEE International Electron Devices Meeting (IEDM) Technical Digest,pp.* 6.7.1 - 6.7.4, 2014.

[90] S. Kim *et al.*, "Latch-up based bidirectional npn selector for bipolar resistance-change," *Appl. Phys. Lett.*, vol. 103, pp. 033505, 2013.

[91] P. Stoliar *et al.*, "Universal Electric-Field-Driven Resistive Transition in Narrow-Gap Mott Insulators," *Advanced Materials*, vol. 25, pp. 3222-3226, 2013.

[92] Q. Luo *et al.* "Cu BEOL Compatible Selector with High Selectivity (>107), Extremely Low Off-current (~pA) and High Endurance (>1010)," *2015 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 253-256, 2015.

[93] D. Kau *et al.*, "A stackable cross point phase change memory," *2009 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 571-574, 2009.

[94] M. Anbarasu, M. Wimmer, G. Bruns, M. Salinga, and M. Wuttig, "Nanosecond threshold switching of $GeTe_6$ cells and their potential as selector devices," *Appl. Phys. Lett.*, vol. 100, pp. 143505-143505-4, 2012.

[95] H. Yang *et al.*, "Novel Selector for High Density Non-Volatile Memory with Ultra-Low Holding Voltage and 107 On/Off Ratio," *2015 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. T130 - T131, 2015.

[96] Q. Luo *et al.*, "Demonstration of 3D Vertical RRAM with Ultra Low-leakage, High-selectivity and Self-compliance Memory Cells," *2015 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 245-248, 2015.

[97] M. Son *et al.*, "Excellent Selector Characteristics of Nanoscale VO_2 for High-Density Bipolar ReRAM Applications," *IEEE Electron Device Lett.*, vol. 32, pp. 1579-1581, 2011.

[98] S. Kim *et al.*, "Ultrathin (<10nm) Nb 2O 5/NbO 2 hybrid memory with both memory and selector characteristics for high density 3D vertically stackable RRAM applications," *2012 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. 155-6, 2012.

[99] H. Y. Peng *et al.*, "Deterministic conversion between memory and threshold resistive switching via tuning the strong electron correlation," *Scientific Reports*, vol. 2, pp. 442/1-6, 2012.

[100] S. H. Chang *et al.*, "Occurrence of Both Unipolar Memory and Threshold Resistance Switching in a NiO Film," *Phys. Rev. Lett.*, vol. 102, pp. 26801/1-, 2009.

[101] T. Liu, M. Verma, Y. Kang, and M. Orlowski, "Volatile resistive switching in Cu/TaOx/delta-Cu/Pt devices," *Appl. Phys. Lett.*, vol. 101, pp. 073510, 2012.

[102] B. Govoreanu *et al.*, "a-VMCO: a novel forming-free, self-rectifying, analog memory cell," *2015 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. T132-T133, 2015.

[103] X. Liu *et al.*, "Co-Occurrence of Threshold Switching and Memory Switching in Pt/NbOx/Pt Cells for Crosspoint Memory Applications," *IEEE Electron Device Lett.*, vol. 33, pp. 236-238, 2012.

[104] S.-G. Park *et al.*, "A Non-Linear ReRAM Cell with sub-1μA Ultralow Operating Current for High Density Vertical Resistive Memory (VRRAM)," *201X IEEE International Electron Devices Meeting (IEDM) Technical Digest,pp.501-504*, 2012.

[105] J. J. Yang *et al.*, "Engineering nonlinearity into memristors for passive crossbar applications," *Appl. Phys. Lett.*, vol. 100, pp. 113501/1-, 2012.

[106] J. Woo *et al.*, "Selector-less RRAM with non-linearity of device for cross-point array applications," *Microelectronic Engineering*, vol. 109, pp. 360-363, 2013.

[107] H.D. Lee *et al.*, "Integration of 4F2 selector-less crossbar array 2Mb ReRAM based on transition metal oxides for high density memory applications," *2012 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp. 151-152, 2012.

[108] B. Govoreanu *et al.*, "Vacancy-Modulated Conductive Oxide Resistive RAM (VMCO-RRAM):An Area-Scalable Switching Current, Self-Compliant, Highly Nonlinear and Wide On/Off-Window Resistive Switching Cell," *2013 IEEE International Electron Devices Meeting (IEDM) Technical Digest*, vol. 13, pp. 256-259, 2013.

[109] R. Meyer *et al.*,"Oxide Dual-Layer Memory Element for Scalable Non-Volatile Cross-Point Memory Technology," *Proceedings 2008 9th Non-Volatile Memory Technology Symposium (NVMTS), p*p. 54-58, 2008.

[110] C. J. Chevallier *et al.*, "A 0.13μm 64Mb Multi-Layered Conductive Metal-Oxide Memory," *2010 IEEE International Solid-State Circuits Conference (ISSCC) Technical Digest*, pp. 260, 2010.

[111] M. Jo *et al.*,"Novel Cross-point Resistive Switching Memory with Self-formed Schottky Barrier," *2010 Symposium on VLSI Technology (VLSIT) Digest of Technical Papers*, pp.53-54, 2010.

[112] X. Liu *et al.*, "Complementary resistive switching in niobium oxide-based resistive memory devices," *IEEE Electron Device Lett.*, vol. 34, pp. 235-237, 2013.

[113] Y. Yang, P. Sheridan, and W. Lu, "Complementary resistive switching in tantalum oxide-based resistive memory devices," *Appl. Phys. Lett.*, vol. 100, pp. 203112/1-4, 2012.

[114] Y. Yang, S. Choi, and W. Lu, "Oxide Heterostructure Resistive Memory," *Nano Letters*, vol. 13, pp. 2908-2915, 2013.

[115] G. Tang *et al.*, "Programmable complementary resistive switching behaviours of a plasma-oxidised titanium oxide nanolayer," *Nanoscale (UK)*, vol. 5, pp. 422-428, 2013.

E4 From Memristive Gate-Array Logic to Neuromorphic Computing

Eike Linn and Arne Heittmann
RWTH Aachen University, Germany

Contents

1 Introduction **2**

2 Memristors and Dynamical Systems **2**
2.1 Two terminal devices 2
2.2 Dynamical Systems 3
2.3 Memristive Modeling Evaluation criteria 5
2.4 Linear memristor models 7
2.5 A TaOx-based non-linear memristive model 13

3 ReRAM-based Logic Approaches **16**
3.1 Different ReRAM-based approaches 16
3.2 In array computation 20

4 ReRAM-based Neuromorphic Circuits **24**
4.1 Artificial Neural Networks 24
4.2 CMOS neuromorphic circuits 28
4.3 ReRAMs for neuromorphic circuits 32

Lecture Notes of the 47th IFF Spring School "Memristive Phenomena – From Fundamental Physics to Neuromorphic Computing" (Forschungszentrum Jülich, 2016).

1 Introduction

These lecture notes cover three aspects: In section 2 the required definitions for memristive systems and memristors are given, and corresponding modeling approaches are discussed. In section 3 the feasibility of logic-in-memory using ReRAM arrays is highlighted. Finally, in section 4 ReRAM properties making these devices highly attractive for neuromorphic computing are discussed and basic constraints are given.

These lecture notes are mostly based on texts and figures of the following publications: Phd thesis of E. Linn [1], book chapters 2 and 25 of [2], and the following scientific papers: [3-9]. The texts are revised and reassembled to fit the outline of the underlying lecture. The original source of each figure is stated in the captions.

2 Memristors and Dynamical Systems

The modelling of emerging devices is an important task to enable proper circuit design and architecture development. Two-terminal devices such as ReRAM cells can be viewed as lumped elements that can be modeled either by ideal circuit elements, or as a dynamical system. In this respect, the classification of ReRAM devices as 'memristor' or 'memristive system' is critically reviewed. The general importance of an in-depth physical insight into the device behavior in order to describe ReRAM devices accurately is elaborated on and important evaluation criteria are introduced.

2.1 Two terminal devices

Electronic devices can be characterized by the number of terminals. For example a MOSFET offers four terminals (Source, Drain, Gate, Bulk) while a *pn*-Diode offers two terminals like ReRAM cells. Moreover, two terminal device models often neglect the spatial dimension of the device, thus are considered lumped elements.

Lumped Elements

In electrical engineering, devices are in general considered as lumped elements. The lumped element assumption is valid if the device size is $\ll \lambda/4$ (wavelength of the excitation signal). A basic feature of lumped elements is the time-invariance, i.e. there is no explicit dependency on t in the device model.

Ideal Circuit Element Approach

There are three basic linear circuit elements: resistor R, capacitance C and inductor L.

According to Chua, also ideal non-linear circuit elements can be defined via the variables V, I, ϕ and q, resulting in four non-linear passive devices: non-linear resistor, non-linear capacitor, non-linear inductor and (ideal) memristor (M). Even higher order circuit elements are feasible following this line [10].

One can use this ideal circuit elements for device modelling, especially black box modelling, resulting in an abstract circuit representing the device. See for example [11].

2.2 Dynamical Systems

Instead of using a number of ideal circuit elements to model a device, a system of differential equations describing the device behaviour can be used for modelling. Such a dynamical system is then implemented for example in SPICE and used for simulations. Since the dynamical system should reflect the actual device physics in detail, the obtained simulation results are more general compared to basic device simulations using predefined threshold voltages.

Introduction of Dynamical Systems

A dynamical system can be represented by two equations:

$$\begin{aligned} \boldsymbol{y} &= h(\boldsymbol{x}, \boldsymbol{u}, t) \\ \dot{\boldsymbol{x}} &= f(\boldsymbol{x}, \boldsymbol{u}, t) \end{aligned} \tag{1}$$

The first equation is the readout equation and the second is the state equation [12]. The functions $f(\cdot)$ and $h(\cdot)$ are non-linear and dependent on time t as well as on variables $\boldsymbol{x}$ and $\boldsymbol{u}$, which are multidimensional in general. The variable $\boldsymbol{x}$ stands for state of the system, the input variable $\boldsymbol{u}$ reflects external excitations, and $\boldsymbol{y}$ is the output or observation variable. This system-theoretical approach can be applied to arbitrary systems and is widely used in electrical engineering.

Memristive Systems

A memristive system [13] is a special case of dynamical system (eq. (1)) and displays a generic term for a complete class of two terminal devices. These devices are characterized by a so-called 'pinched hysteresis loop' (see Figure **1**, [14, 15]) and are non-linear in general. A memristive system is defined by the state-dependent Ohm's law and the state equation, which is multi-dimensional and time-dependent in general. Note that a memristor [16] is only a special case of a first-order memristive system where the state variable equals the flown charge ($x = q$).

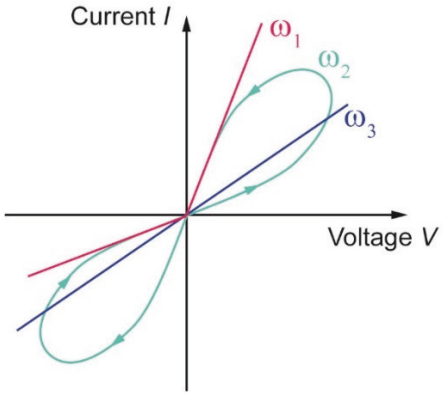


Figure 1: *A pinched hysteresis I-V-loop is the identifying feature of a memristive system. For intermediate frequencies (e.g., ω_2) such a loop is visible. For very low frequencies ($\omega_1 \rightarrow 0$) a memristive system is indistinguishable from a non-linear resistance, while for very large frequencies ($\omega_3 \rightarrow \infty$) a memristive system is reduced to a linear resistance. For all frequencies the curves are pinched, which means that all curves run through the origin (0,0). From [1].*

The relevance of the memristive system approach results from the possibility to model dynamic device behaviour of resistive switching elements. This was first recognized by the HP group in 2008 [17] and has triggered many groups to work on memristive circuit models [14].

For a memristive system, y and u are scalar values and y is a product of h and u:

$$\begin{aligned} y &= h(\boldsymbol{x}, u, t) \cdot u \\ \dot{\boldsymbol{x}} &= f(\boldsymbol{x}, u, t) \end{aligned} \tag{2}$$

Equation (2) shows that y is always zero when u is zero, which corresponds to a Lissajous figure with a pinched hysteresis loop (Figure **1**).

The definition of memristive systems given in [13] can be directly applied on two terminal electronic devices. Memristive systems can be either current controlled ($u = I$ and $h = R$) or voltage controlled ($u = V$ and $h = G$). In this case, u is the input variable, while y is the output variable. A current controlled memristive system reads

$$\begin{aligned} V &= R(\boldsymbol{x}, I, t) \cdot V \\ \dot{\boldsymbol{x}} &= f(\boldsymbol{x}, I, t), \end{aligned} \tag{3}$$

whereas a voltage controlled memristive system reads

$$\begin{aligned} I &= G(\boldsymbol{x}, V, t) \cdot I \\ \dot{\boldsymbol{x}} &= f(\boldsymbol{x}, V, t). \end{aligned} \tag{4}$$

Time-Invariant Memristive Systems

A memristive system is considered time-invariant when neither f nor R (or G, respectively) is time-dependent. This limitation leads to the more common description of memristive systems. A current controlled memristive system then reads

$$\begin{aligned} V &= R(\boldsymbol{x}, I) \cdot V \\ \dot{\boldsymbol{x}} &= f(\boldsymbol{x}, I). \end{aligned} \tag{5}$$

The corresponding voltage controlled memristive system reads

$$\begin{aligned} I &= G(\boldsymbol{x}, V) \cdot I \\ \dot{\boldsymbol{x}} &= f(\boldsymbol{x}, V). \end{aligned} \tag{6}$$

For modelling of bipolar resistive switches the time-invariant formulation is sufficient. Note that there are two additional conditions which must be fulfilled. To result in a pinched hysteresis loop

$$R(\boldsymbol{x}, 0) \cdot I = 0 \tag{7}$$

and accordingly

$$G(\boldsymbol{x}, 0) \cdot V = 0 \tag{8}$$

must hold. Additionally, for a non-volatile memory device

$$f(\boldsymbol{x}, 0) = 0 \tag{9}$$

must hold true, because no change of state should occur without external excitation. For modelling of resistive switches as a memristive system it is crucial to identify inner state variables. At least one state variable describing a structural change is needed, e.g., the length of a filament in electro-chemical metallization cells (ECM).

Simple Time-Invariant Memristive Systems

In the simplest case the resistance R or conductance G, respectively, are only functions of the state variable $\boldsymbol{x}$ and, $\dot{\boldsymbol{x}}$ is only a function of current I or voltage V, respectively. The current controlled case is defined as

$$\begin{aligned} V &= R(\boldsymbol{x}) \cdot V \\ \dot{\boldsymbol{x}} &= f(I). \end{aligned} \tag{10}$$

The voltage controlled case reads

$$\begin{aligned} I &= G(\boldsymbol{x}) \cdot I \\ \dot{\boldsymbol{x}} &= f(V). \end{aligned} \tag{11}$$

This modeling approach was suggested in [17] and is used for SPICE implementation in section 2.4 where also the limitations of this approach are highlighted [3].

Memristor

A memristor is sometimes considered the forth passive circuit element [16] and is a special case of a memristive system. A memristor has only one state variable which is the flown charge $x = q$ in the current controlled case

$$\begin{aligned} V &= R(q) \cdot V \\ \dot{q} &= I. \end{aligned} \tag{12}$$

In the voltage controlled case the magnetic flux is the state variable $x = \phi$

$$\begin{aligned} I &= G(\phi) \cdot I \\ \dot{\phi} &= V. \end{aligned} \tag{13}$$

Note: in general, bipolar resistive switches cannot be modelled as ideal memristors.

2.3 Memristive Modeling Evaluation criteria

In memristive modeling a resistive switch is considered a dynamical system (eq. (1)) and it should be possible to simulate the device behaviour for a wide range of input signals. This is the main strength of the memristive modelling approach compared to using models with built-in fixed threshold voltages [18, 19], which are only valid for certain input signals.

In [3] three basic evaluation criteria are introduced:

The **first evaluation criterion** considers the *I-V* characteristics of bipolar resistive switches (Figure 2a) which exhibit some distinct features that should be reproduced by a suitable model (compare [20], for example).

Typically, during the SET operation an abrupt increase in current is observed for ECM and VCM-based ReRAM cells (VCM = Valence Change Mechanism). The RESET operation, however, differs for these two classes: VCM devices often show a gradual RESET whereas ECM devices exhibit an abrupt change. Furthermore, the *I-V* characteristics are asymmetric with respect to the origin. In addition, the SET and RESET voltages increase when the sweep rate of the voltage sweep is increased. Additionally, it is know from experiment that a wide range of excitation signals will lead to resistive switching device behaviour. In so far, a suitable memristive model should offer certain robustness against changes in the input voltage amplitude and variations of initial values, e.g. the initial value of the state variable.

The **second criterion** is related to the switching kinetics. In experiments a strong non-linear relationship between SET time t_{SET} and pulse height V_p is observed (see Figure 3). Here device data from four typical VCM devices is selected: strontium titanate [21], tantalum oxide [22], hafnium oxide [23], and titanium oxide [24]. A common fingerprint of all VCM devices is the decrease of t_{SET} by several orders of magnitude by only increasing the pulse amplitude V_p by a factor of two. Hence, the second criterion is the check for such an exponential dependency. Fulfilling this criterion is essential to enable simulation of typical applications using memristive devices (either memory or logic applications) which are conducted by fast pulses.

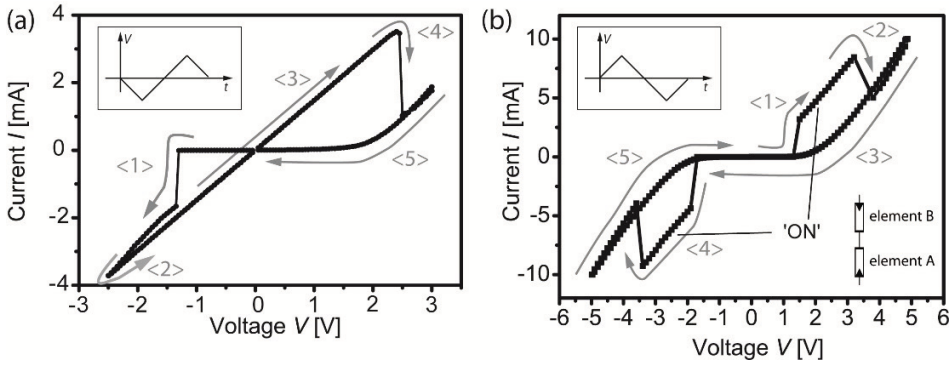


Figure 2: *(a) Exemplary I-V characteristic of a TaO_x-based resistively switching device. The input signal is a triangular voltage sweep of amplitude -2.5 V/ 3V (see inset). Note that for a symmetric voltage amplitude (e.g., -3 V/ 3V) the characteristic does not change significantly. Initially, the device is in HRS state and switches to state LRS at <1>. In <2>, <3> the device stays in the LRS and switches back to HRS at <4>. This is again the initial HRS state. Note that the switching polarity depends on the actual material composition. Here, the input voltage was applied to the top electrode (Pt) while the bottom electrode (Ta) was grounded. (b) I-V characteristic of a TaO_x-based complementary resistive switch device. The input signal is a triangular voltage sweep of amplitude 5 V (see inset). Starting from HRS/LRS (element A/element B), the device switches to LRS/LRS ('ON state') first (<1>). Next, the CRS cell switches over to LRS/HRS (<2>) and remains in this state until the negative SET voltage is reached (<3>). Then, the device switches again to LRS/LRS (<4>), and later on back to HRS/LRS (<5>). From [3].*

A **third criterion** arises from the need for multi-element simulations for mapping real-world circuits. Two-element circuits are the simplest application, and therefore, are well suited for a basic analysis. Here, an anti-serial connection of two memristive devices, known as complementary resistive switch (CRS), is considered [25]. Typical *I-V* characteristics of VCM-type CRS cells can be found in [26], for example (see Figure 2b). One distinctive feature of the anti-serial connection is the presence of an overall low resistive state ('ON state') when applying a voltage sweep. The existence of this ON state region in a two element simulation can be used as a further check for model consistency [27].

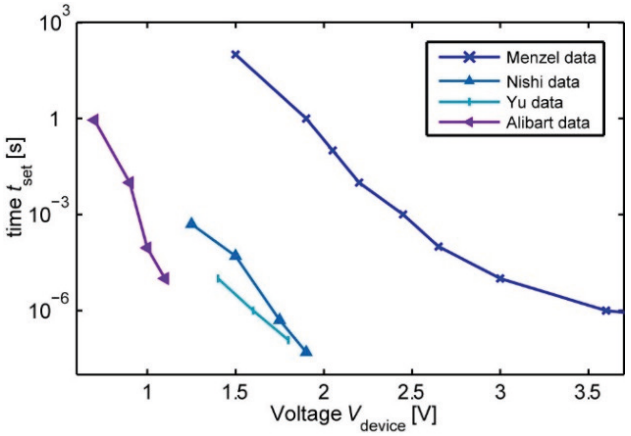


Figure 3: *Set time t_{SET} of the switching from HRS to LRS versus applied pulse height V_p. Menzel data for $SrTiO_x$ [21], Nishi data for TaO_x [22], Yu data for HfO_x [23], and Alibart data for TiO_x [24]. From [3].*

2.4 Linear memristor models

In this section, the basic memristor models are evaluated in terms of the introduced criteria, as presented in [3], starting from the following device equations given in [17]:

$$\dot{x} = h(I) = K_1 \cdot I \tag{14}$$

$$V = R(x) \cdot I = \left(\left(R_{\mathrm{LRS}} - R_{\mathrm{HRS}} \right) \cdot x + R_{\mathrm{HRS}} \right) \cdot I \tag{15}$$

Here, K_1 is a constant, R_{LRS} is the resistance of the low resistive state (LRS) and R_{HRS} is the resistance of the high resistive state (HRS). The state variable x, which represents the position of the boundary between low conductive and high conductive region, is normalized by the switching layer thickness of $D = 10$ nm.

Note that the equation system (14)-(15) is mathematically not a linear system. However, this model is considered a 'linear model' for the following reasons: first, the state variable x influences the resistance $R(x)$ linearly. Second, the voltage V is directly proportional to the current I for a certain $R(x)$. Third, $\dot{x}$ depends linearly on the current I.

To prevent nonphysical values for x, the state variable must be limited to the layer thickness, thus $0 \le x \le 1$ holds. In common SPICE implementations, this bounding is realized by a window function $f(x,I)$. Thus, equation (14) can be rewritten as:

$$\dot{x} = h(x, I) = K_1 \cdot I \cdot f(x, I) \tag{16}$$

Four different window function implementations are considered in the following: Benderli's model [28], Joglekar's model [29], Biolek's model [30] and Shin's model [31], which are illustrated in Figure 4a, c, e and g, respectively.

In case of Biolek's model and Shin's model, the window function also depends on the sign of current I, compare Figure 4e, g. However, one can rewrite Biolek's window function as follows:

$$f(x, I) = \begin{cases} 1 - (x)^{2p} & \text{for } I \ge 0 \\ 1 - (x-1)^{2p} & \text{for } I < 0 \end{cases} \tag{17}$$

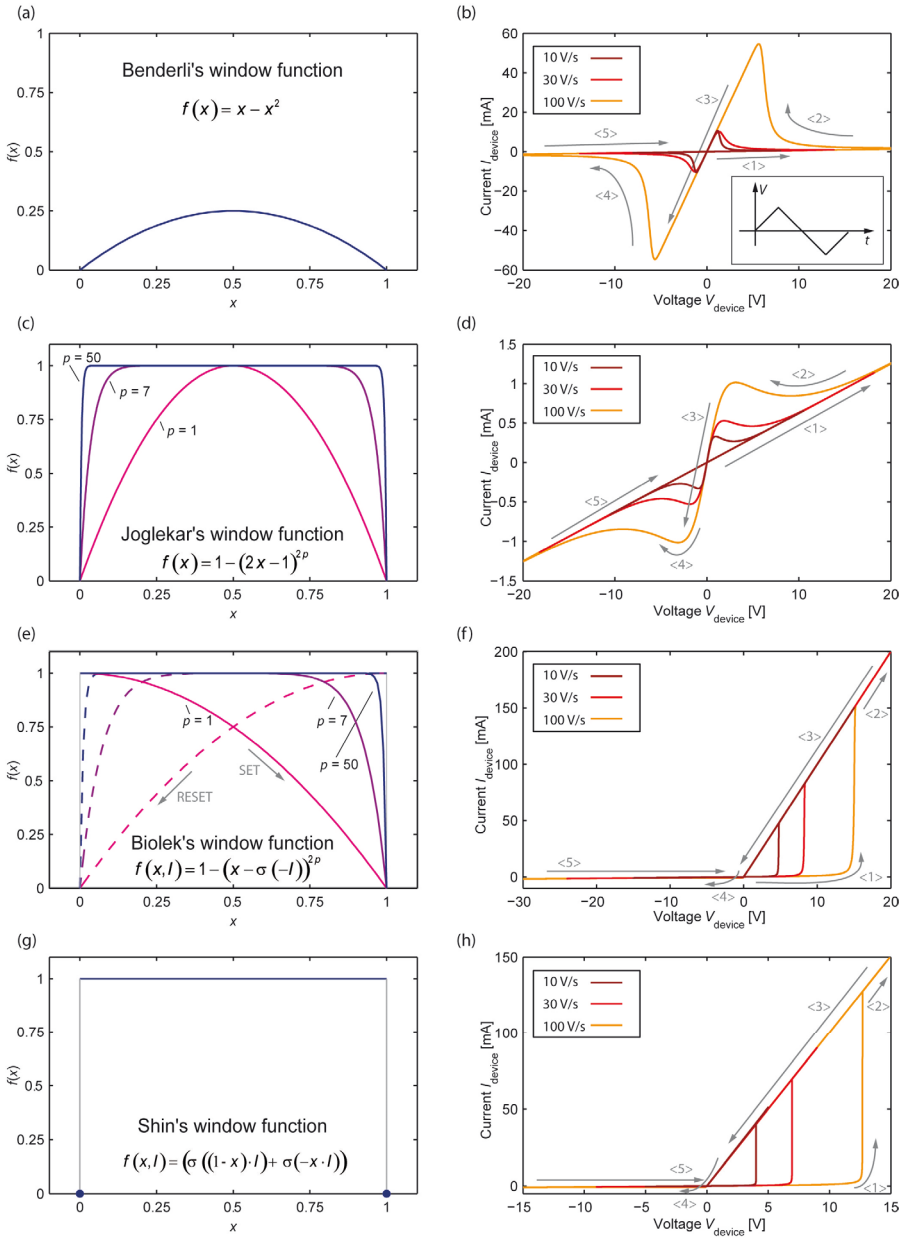


Figure 4: *Implementation of Strukov's model using different window functions. For each model the applied window function and corresponding I-V-characteristics are depicted. (a,b) Benderli's model, (c,d) Joglekar's model, (e,f) Biolek's model and (g,h) Shin's model. The inset in (b) illustrates the input triangular voltage signal. Arrows and numbers <#> indicate the run of the curve. The initial state x_0 for Shin's and Biolek's window was $x_0 = 0$, while for Benderli's window $x_0 = 0.002$ and $x_0 = 10^{-12}$ for Joglekar's window was used. From [3].*

Similarly, Shin's window reads:

$$f(x;I) = \begin{cases} \sigma(1-x) & \text{for } I \geq 0 \\ \sigma(x-1) & \text{for } I < 0 \end{cases} \tag{18}$$

Here, $\sigma(\cdot)$ is the step function. Thus, when only either purely positive or negative input signals are considered, the window function is only a function of x.

The simulation parameters for all models are selected according to [17]: $K_1 = 10^4$ $A^{-1}s^{-1}$, $R_{LRS} = 100\ \Omega$ and $R_{HRS} = 16\ k\Omega$. In Figure 4c Joglekar's window function is shown. It has the same shape as the Benderli's window function, but can be parameterized with p to vary the gradient. To illustrate the impact of p, three curves, for $p = 1$, 7 and 50, are depicted. For the *I-V* simulations in Figure 4d $p = 1$ is applied.

The Biolek's window function offers a similar parameterization as Joglekar's window function, and window curves for $p = 1$, 7 and 50 are shown in Figure 4e. However, the function behaves quite different due to the involved step function $\sigma(\cdot)$ which enables a sudden upward transition of the window function from a value close to zero towards a value close to unity if the sign of the current I changes and when x is in proximity of its limits (see Figure 4e). In Figure 4e the solid line is valid for positive currents (SET direction) and the dashed line for negative currents (RESET direction). Shin's window function (Figure 4g) can be considered an edge case of Biolek's window function for $p \to \infty$.

In order to evaluate these models against the first criterion their *I-V* characteristics is simulated. For this, symmetric triangular input voltage signals with sweep rates of 10 V/s, 30 V/s and 100 V/s (see inset in Figure 4b) are used. The resulting *I-V* characteristics using Benderli's and Joglekar's window function are depicted in Figure 4d, respectively.

Both models exhibit completely symmetric *I-V* characteristic with respect to the origin due to their symmetric window functions, i.e., one switching event (SET) occurs after reaching the maximum voltage level (<2> in Figure 4b and Figure 4d) and the other one (RESET) occurs before reaching the maximum absolute voltage level (<4> in Figure 4b and Figure 4d). Therefore, the symmetry is an inherent property of these two models and as a consequence they cannot reproduce the asymmetry of bipolar resistive switching, as analytically proved [20]. Keep in mind that Benderli's and Joglekar's model represent memristors with window depending on state only. Fingerprint of these types of models is a symmetric *I-V* characteristic with respect to the origin (compare [32, 33]), i.e. criterion 1 is not hold. Furthermore, Benderli's and Joglekar's model do not show an abrupt SET transition which limits their applicability (cf. Figure 4b, and d). In contrast, the simulated *I-V* curves using Biolek's and Shin's window function show an abrupt SET transition. In addition, the *I-V* characteristics are asymmetrical with respect to the origin. However, one should note that the characteristics in Figure 4f, h differ strongly from characteristics of typical VCM devices [20] anyway: for example, the current in LRS for negative voltages is very low.

All models offer the general trend of higher SET voltages for increasing sweep rates. The window functions offer quite different robustness with respect to input signals. Both the Benderli's model and the Joglekar's model tend to stick at the boundary if voltage amplitudes become larger. This is due to the form of the window function which gives rise to convergence problems (called "terminal-state problem" in [34]), and limits applicability of the model. To enable proper simulation, parameters and input signal must be carefully adjusted for these window functions. This problem was solved in the Biolek and Shin model by resetting the window function when the sign of the input signal changes.

In the following, the models are evaluated with respect to the reproducibility of the experimentally observed switching kinetics, i.e. the second evaluation criterion. Thus, voltage pulses of height $V_{device} = V_p$ are applied to each model being initially in the high resistive state (HRS). For positive voltages the models are SET to the low resistive state (LRS) after the time t_{SET}.

From Figure 5a, one can see that the basic trend observed in experiments, i.e., decreasing t_{set} for increasing pulse height, is also observed for the four simulation models. But, the dependency is much less pronounced as observed in experiments. Compared to experimental data shown in [24] for titanium oxide, one can clearly see that actual dependency differs greatly (several orders of magnitude).

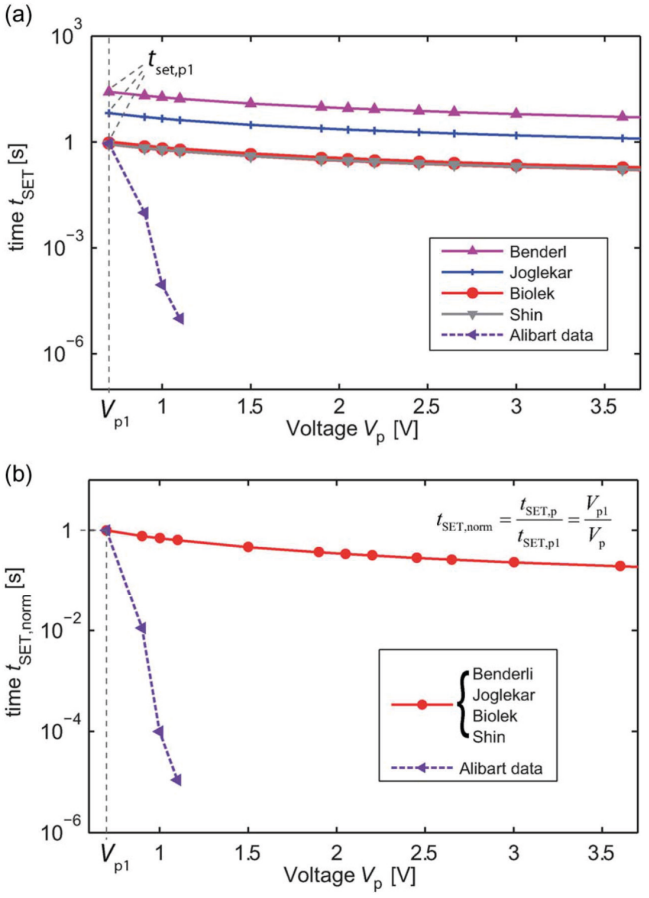


Figure 5: *Set time t_{SET} of the switching from HRS to LRS versus applied pulse height V_p. In simulations t_{SET} is defined at x = 0.5. (a) shows the raw data. (b) depicts normalized data with respect to the V_{p1} = 0.7 V points. From [3].*

This mismatch can be directly assigned to the *R-x* dependency in model equation (4), which is not sufficiently non-linear. Furthermore, it was shown that independent of the applied window function, all models offer the same kinetics (cf. Figure 5b). By normalizing the t_{SET} values by a certain point (here: V_{p1} = 0.7 V), all curves collapse to a single line, showing this feature directly. By inserting eq. (14) into eq. (15) results in:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = K_1 \cdot \frac{V(t)}{R(x)} \cdot f(x) \tag{19}$$

where the dependence of $f(\cdot)$ on I is removed, since for $t \in [t_0, t_1]$ the current has a unique sign and thus $f(\cdot)$ depends only on x. Equation (19) is a differential equation offering two variables, x and t. Integration by parts results in:

$$\int_{x_0}^{x_1} \left(\left(R_{\mathrm{LRS}} - R_{\mathrm{HRS}} \right) \cdot x + R_{\mathrm{HRS}} \right) \frac{1}{K_1 \cdot f(x)} \cdot \mathrm{d}x = \int_{t_0}^{t_1} V(t) \cdot \mathrm{d}t \tag{20}$$

The bounds of integration are: $t_0 = 0$ s and $t_1 = t_{\mathrm{set,p}}$ for the right side and $x_0 = 0$ and $x_1 = 0.5$ (assumed SET condition) for the left side of equation (20). Furthermore, a voltage pulse of amplitude V_{p}, thus $V(t) = V_{\mathrm{p}}$ holds, is considered. Finally, the equation reads:

$$\int_{0}^{0.5} \left(\left(R_{\mathrm{LRS}} - R_{\mathrm{HRS}} \right) \cdot x + R_{\mathrm{HRS}} \right) \frac{1}{K_1 \cdot f(x)} \cdot \mathrm{d}x = V_{\mathrm{p}} \int_{0}^{t_{\mathrm{SET,p}}} \mathrm{d}t \tag{21}$$

which is equivalent to

$$K_2 = V_{\mathrm{p}} \cdot t_{\mathrm{SET,p}} \tag{22}$$

The left hand side of the equation (21) is constant for every model and is called K_2 in equation (22). Note that the value of K_2 is specific to the applied model, but cancels out when normalizing values with respect to a certain pulse height V_{p1} offering a set time $t_{\mathrm{SET,p1}}$. This procedure is done for each model independently, and results in the graph shown in Figure 5b. Thus, for all models of this kind the normalized SET time only depends on the pulse height V_{p} and not on the window function (Figure 5b):

$$t_{\mathrm{SET,norm}} = \frac{t_{\mathrm{SET,p}}}{t_{\mathrm{SET,p1}}} = \frac{V_{\mathrm{p1}}}{V_{\mathrm{p}}} \tag{23}$$

The resulting dependency is

$$t_{\mathrm{SET}} \sim \frac{1}{V_{\mathrm{p}}} \tag{24}$$

So, the models are not capable to show the required exponential dependency.

From these considerations it is clear that a simple addition of a window function is not appropriate to introduce realistic device dynamics to the initial memristor model.

Next the third criterion which is the anti-serial connection of two elements. Here, the results reveal even more striking mismatches between simulation and real device behaviour. Due to the anti-serial connection of both cells A and B $\dot{x}_{\mathrm{A}} = -\dot{x}_{\mathrm{B}}$ holds if $f(x_{\mathrm{A}}) = f(x_{\mathrm{B}})$, where the dependence on I is dropped. Therefore, any change of state variable in cell A is cancelled out by the change of state in cell B. Thus, the total resistance of both elements is constant all the time (see Figure 6a) which is not the case in reality, as it is known from CRS cells (compare Figure 2b, where $x_{0\mathrm{A}} \approx x_{0\mathrm{B}} \approx 0$, i.e. $x_{\mathrm{A}}(t = 0) \approx 0$ and $x_{\mathrm{B}}(t = 0) \approx 1$). For Shin's window function this property has been shown in [35]. Biolek's window (Figure 4e) shows the same behaviour (compare Figure 4a) while Benderli's window and the Joglekar's window only offer a straight line for symmetrical initial conditions, e.g., $x_{0\mathrm{A}} = x_{0\mathrm{B}} = 0.001$ ($x_{\mathrm{A}}(t = 0) = x_{0\mathrm{A}}$ and $x_{\mathrm{B}}(t = 0) = 1 - x_{0\mathrm{B}}$).

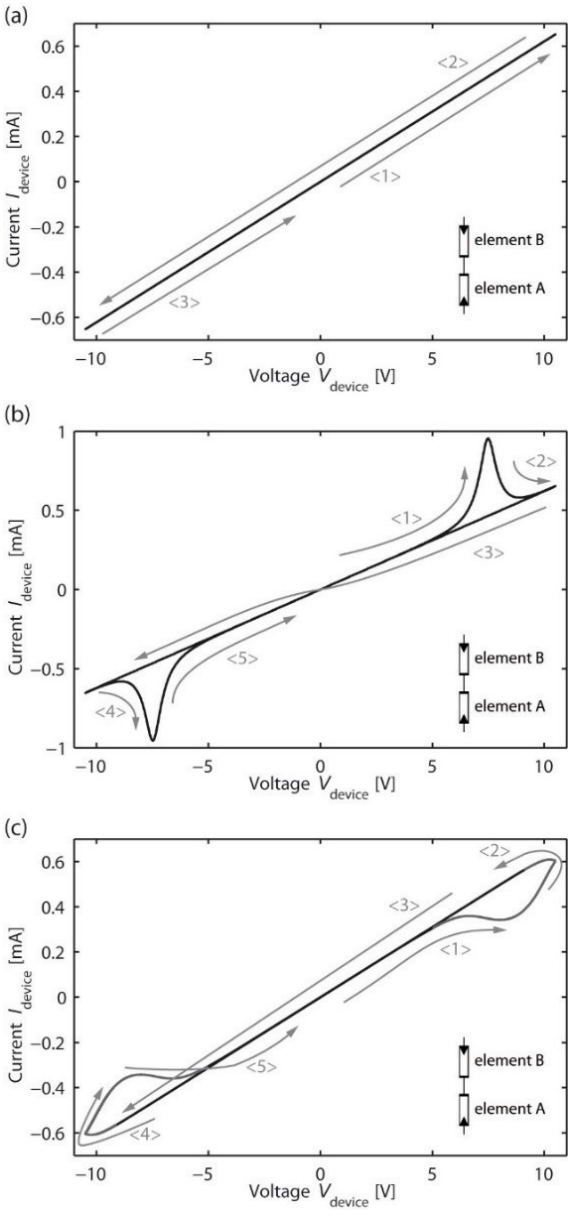


Figure 6: *Anti-serial connected device model simulations using Joglekar's window. A triangular input voltage signal of sweep rates of 10 V/s was used. The curves run-through is denoted by the arrows.*
(a) For symmetrical initial conditions ($x_{0A} = x_{0B} = 0.001$), no change of the overall resistance is observed.
(b) Simulation for $x_{0A} > x_{0B}$ ($x_{0A} = 0.001$ and $x_{0B} = 0.0001$).
(c) Simulation for $x_{0A} < x_{0B}$ ($x_{0A} = 0.0001$ and $x_{0B} = 0.001$). The portion of the loop relative to the regime of increased resistance is marked by a grey line color. From [3].

For Benderli's window and the Joglekar's window one can force a non-ohmic device behaviour by starting from different initial states and considering a smooth window function. For simulation shown in Figure 6b the model with Joglekar's window ($p = 1$) is used and $x_{0A} > x_{0B}$ is assumed. Due to the asymmetry of the initial values the SET process in element A starts earlier, leading to an increased current <1> in Figure 6b. Note that a similar result was observed in [36] which is in

accordance to CRS behaviour at first glance. However, for the negative voltage cycle the increased current (point <4>) occurs after reaching the maximum absolute voltage which does not correspond to real device behaviour at all. For $x_{0A} < x_{0B}$ the observed behaviour becomes even more unusual since the resistance (chordal resistance) is increased in a certain regime (grey line curves at points <1>, <4> in Figure 6c) – the opposite behaviour than observed in experiments. In consequence, the simulation result strongly depends on the initial states, which is unfavourable according to the first criterion. However, adjusting the initial states is not suited to reproduce real device behaviour for those models. Pay attention that a strong dependency on initial states is a commonly observed incident for memristor models (compare e.g. [37]).

In conclusion, although highly attractive due to ease of use, none of the above studied models is suited to reproduce the basic resistive switch properties for arbitrary input signals. But these models could be modified to meet the requirements. The window function can have a positive impact on the accuracy, but cannot fix the basic physical equations. Correspondingly, simulation results obtained from these models, e.g. [38-40], should be reconsidered using more sophisticated models.

However, one should keep in mind that additional model complexity allows a very high predictivity for a variation of parameter inputs. But this complexity might also give rise to convergence issues.

2.5 A TaOx-based non-linear memristive model

As an example, of a physics-based memristive model a Pt/TaO_x/Ta model is discussed [4]. This model is based on Hur's model [41], but considering the interface Pt/TaO_x as a Schottky barrier with barrier lowering effects. In contrast, to Hur's approach the Schottky diode is considered also in the low resistive state (LRS). As basic structure a plug/disc structure (compare [21]) is assumed, where the well conducting plug is modelled as a series resistor (Figure 7). Switching takes place only in the disc region of fixed length L_{Disc}. The average oxygen vacancy concentration N in the disc is considered as the state variable which varies the electronic resistance R_{el} and the effective Schottky barrier height. It is important to note that one distinguishes between the ionic current I_{Ion} (which causes the change of state variable) and the electronic current (which represents the measureable device current). To enable a comparison of simulation results to experimental CRS data [26], this model is fitted to a L_{TaOx} = 11 nm Pt/TaO_x/Ta device *I-V* characteristic measured by Nishi and Schmelzer ([22], Fig. 2b), see Figure 8. Note that for this device thickness a quite large area dependent parallel current (I_{Area}) is present, which is the dominant current share in the high resistive state (HRS) [22]. Parameter details can be found in [4].

For simulation a VerilogA behavioural model of the single Pt/TaO_x/Ta device was implemented. This module can then be connected to other circuit elements, e.g. a second module with reversed pins to form a complementary resistive switch. As can be seen in Figure 8, the criterion 1 (basic *I-V* non-symmetric) is fulfilled by this model. For a next qualitative check of model consistency a triangular input signals of different sweep rates is applied to the memristive model (Figure 9).

As know from experiment, SET and RESET voltages increase with increasing sweep rate. However, the actual device kinetic is extracted by applying voltage pulses of different pulse height to the device model. The resulting SET time (t_{SET}) dependency is depicted in Figure 10. This plot reveals the exponential nature of t_{SET} which is a general property of resistive switches (criterion 2). Note that the quasi-static SET time for the underlying device is relatively large, thus the curve is shifted in parallel compared to the measurement data which reflect the medians of a large set of measurements.

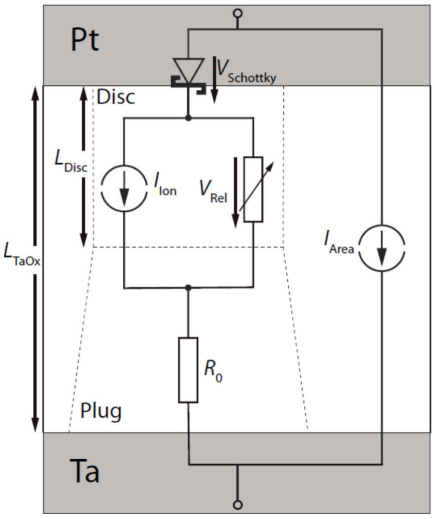


Figure 7: *Eqivalent circuit model of the Pt/TaO$_x$/Ta device. From [4].*

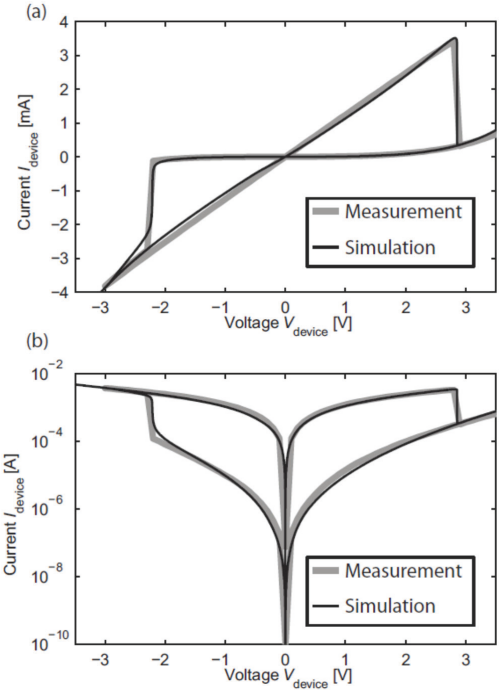


Figure 8: *Measured I-V characteristic of a Pt/TaO$_x$/Ta device [14] and corresponding simulation of the memristive VerilogA model. (a) On linear scale, and (b) on logarithmic scale. From [4].*

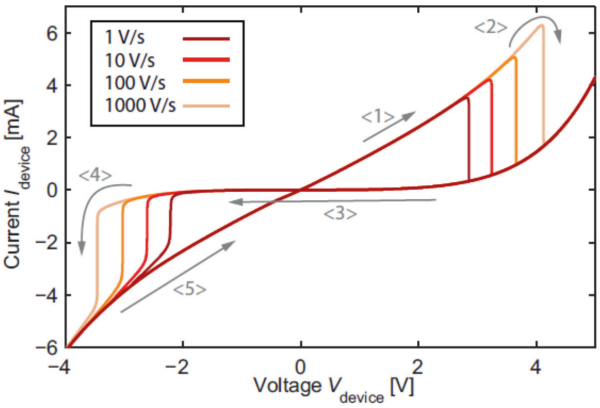


Figure 9: *Sweep rate dependency of the simulated I-V curve of the Pt/TaO$_x$/Ta device. The arrows indicate the run of the curve. From [4].*

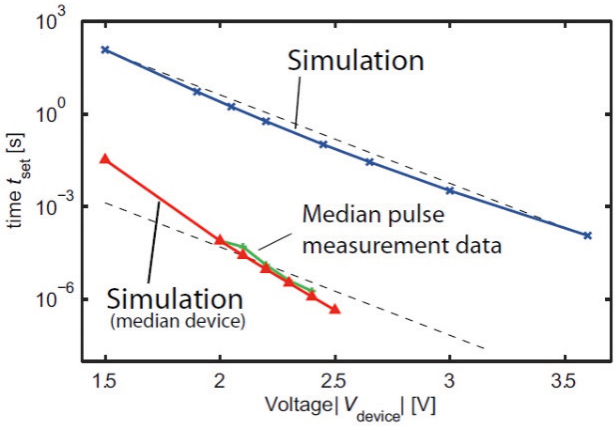


Figure 10: *Pulse height versus SET time for a Pt/TaOx/Ta device (quasi-static $V_{SET} \approx 2.2$ V) and a median device for comparison (quasi-static $V_{SET} \approx 1.4$ V). The dashed trend lines are parallel. From [4].*

In addition, also two anti-serially connected devices, i.e. a complementary resistive switch were simulated. In Figure 11, the resulting *I-V* curve is depicted in linear and logarithmic scale, reproducing very well the shape known for TaO_x CRS devices [13, 6]. In result, also criterion 3 is valid for this memristive model.

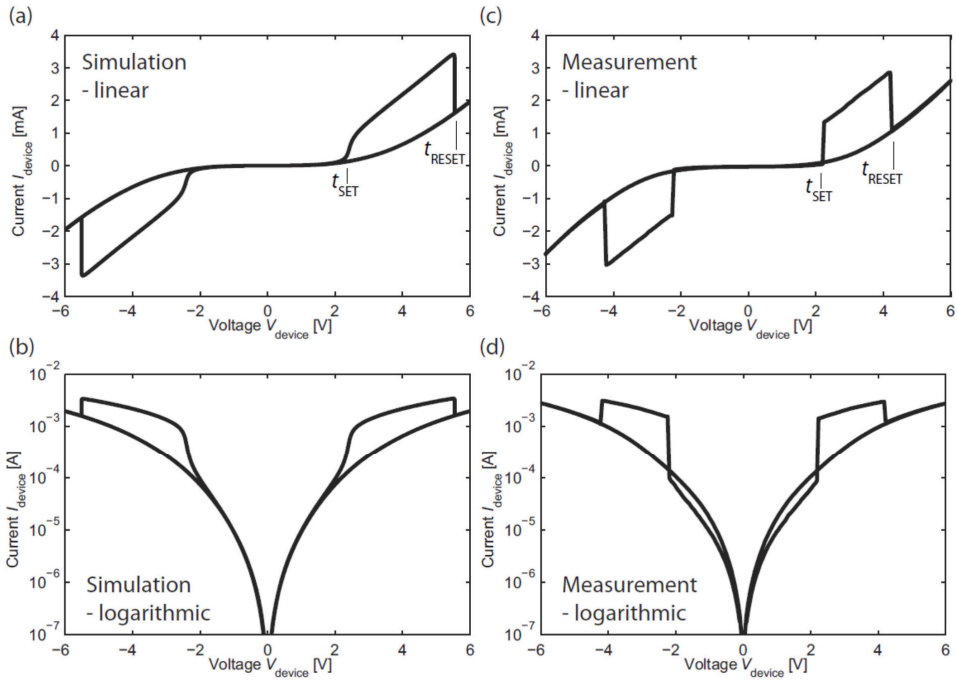


Figure 11: *CRS simulation, (a) linear scale, (b) logarithmic scale. (c) and (d) depict the I-V curves of a typical TaOx-based CRS cell. From [4].*

3 ReRAM-based Logic Approaches

3.1 Different ReRAM-based approaches

There are several ideas how to use resistive switches for reconfigurable logic applications in hybrid CMOS - nanoelectronic circuits. Most concepts aim on realizing field programmable logic arrays (FPGA) or programmable logic arrays (PLA) by use of resistive switches. Concepts can be classified by their mode of application:

- Resistive Switches as Programmable Interconnects
- Resistive Switches as Memory Cell
- Resistive Switches as Latching Device

First concepts for resistive switch-based logic rely on the idea of using programmable interconnects. Since resistive switches are either high- or low-resistive, connections between two lines can be programmed by such non-volatile switches.

In the Teramac concept [42], each memory cell of a conventional memory array controls a resistive switch in a crossbar array, which is processed on top of the memory array (see **Figur 12**). Since one memory cell is needed for each resistive switch, this concept is not very efficient. In fact, because a CMOS-based conventional memory array is essential for the Teramac concept, no benefit compared to purely CMOS-based memory used as look-up table results from this approach [43].

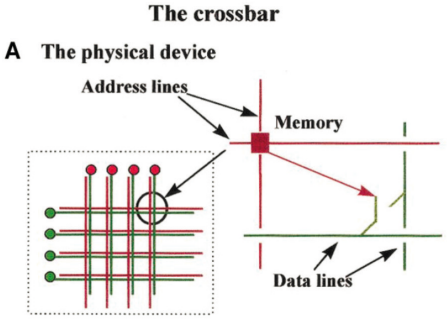


Figure 12: *The Teramac concept. Each resistive crossbar junction is controlled by a memory element. From [42].*

PLA concepts are based on two crossbar arrays of programmable interconnects, one implementing AND to form the minterms and one implementing OR to realize all logic functions in a two level logic representation. PLAs can be used as logic blocks in FPGAs [44], but since logic blocks (and therefore the crossbar arrays) are typically small for realistic FPGA applications [45], CMOS overhead is large. On the other hand, due to the sneak path problem, the size of usable crossbar arrays is limited anyway. A resistive PLA logic block realizing a crossbar full adder is given in [46] (see Figure 13).

There are also PLA concepts with a latching device for storage and signal regeneration integrated in the crossbar array. In [47], the use of two tunnelling diodes, so-called goto pairs, was suggested as latch, while a BRS and a diode were suggested in [48] to form a crossbar latch. For these approaches, two clock signals are needed.

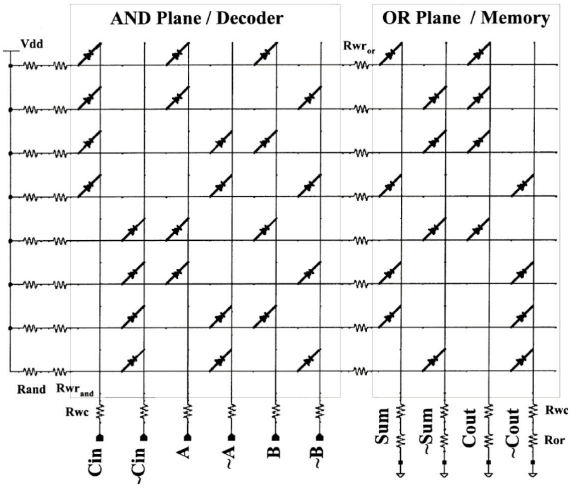


Figure 13: *Resistive PLA crossbar full adder. From [46].*

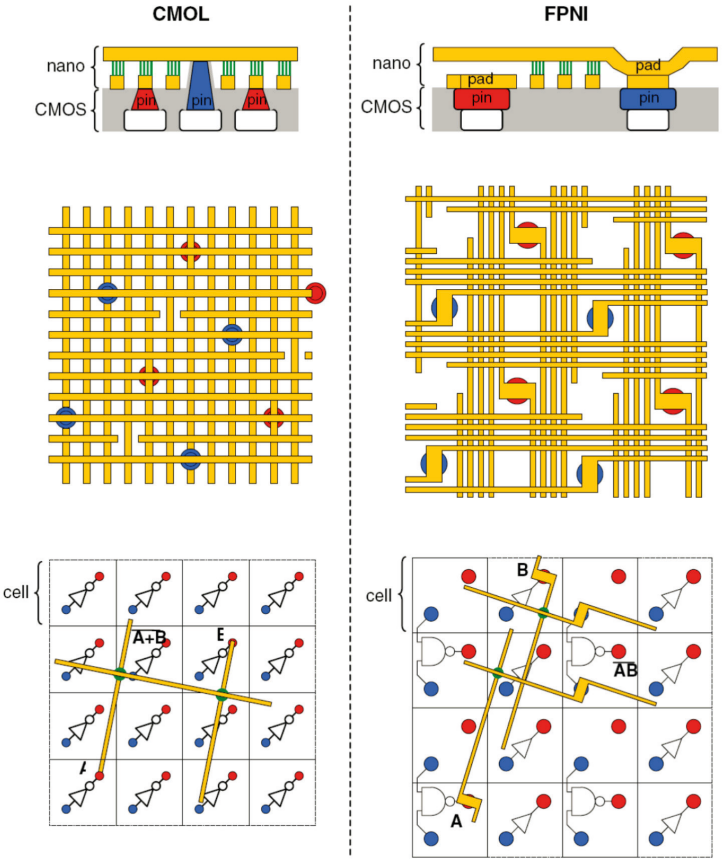


Figure 14: *CMOL and FPNI concept. From [49]*

In the CMOL FPGA concept [43], a sea of elementary CMOS cells, each consisting of two pass transistors and an inverter, is connected to a nano-crossbar array consisting of discontinuous lines (Figure 14). The elementary CMOS cells are connected to each other by programmed (BRS switched to LRS) junctions allowing for wired-or logic. Both nano crossbar layers must be connected to CMOS via nano pins, making fabrication very difficult. Since actual nanowire structure and connectivity must be evaluated after fabrication, mapping is very challenging [49].

A similar concept to CMOL is called FPNI (field programmable nanowire interconnect, see Figure 14). Nano junctions are only used for routing and only one height of nano pins is needed, but crossbar array in FPNI is sparser, degrading performance to about 50% [50].

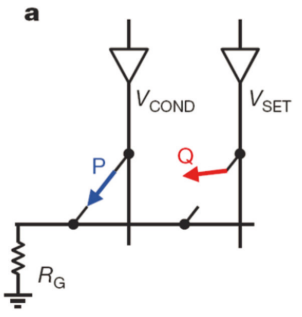
In conclusion, the common feature of these approaches is that resistive switches are configured once, or very infrequent, to adjust a logic function. In consequence, junctions are either set as a closed connection (LRS) or an open connection (HRS), making no use of the inherent memory feature of resistive switches.

A completely different approach is based on memories where, for example, resistive memories are used as look-up tables in FPGAs. In [51], 1T1R memories are suggested as replacements of SRAM-based look-up tables (LUTs). Also, crossbar memory-based LUTs are thinkable, but the overhead is large due to small array sizes used in conventional FPGA design. In [52], another memory-based computing approach for FPGAs with need for large crossbar arrays - and thus small CMOS overhead - is suggested. In this approach, multi-input-multi-output LUTs are mapped on a large crossbar array memory simplifying routing constraints [52, 53].

In full sequential logic concepts, no combinational logic blocks are present. In [54], the (material) implication is given as a basic logic function in need of two BRS and a load resistor R_G forming the 'IMP-gate' (see Figure 15 and compare latch described in [55, 56]). This operation can be performed in four steps (cf. Figure 16a):

1. Set device P to p ($V_P = \pm V_{Write}$)
2. Set device Q to q ($V_Q = \pm V_{Write}$)
3. $q' = p$ IMP q ($V_P = V_{COND}$ and $V_Q = V_{write}$)
4. Read q'

Note that the load resistor must be in the range of $R_{LRS} < R_G < R_{HRS}$.



b

$q' \leftarrow p\text{IMP}q$

In p	In q	Out q'
0	0	1
0	1	1
1	0	0
1	1	1

Figure 15: *IMP operation realized by two bipolar resistive switches and a load resistor in four steps. Reproduced with permission from [54].*

Since IMP and FALSE form a computationally complete logic class, more complex functions such as NAND can also be provided by three BRS and a load resistor in six sequential steps in [54].

Due to the sneak path problem, this concept is limited to word structures or very small arrays, but can be used in an optimized form for CRS cells. Additionally, IMP or latch functionality is an intrinsic feature of a single BRS, thus the number of needed cells can be reduced in so called CRS-logic.

In CRS-logic (approach 2), the input signals $V_p = \pm\frac{1}{2}V_{Write}$ and $V_q = \pm\frac{1}{2}V_{Write}$ are applied at the terminals T1 and T2 of the two terminal device. The final result is stored as resistive state Z. For $Z = p$ IMP q the following steps are performed (Figure 16b):

1. Init device Z to '1' ($V_{T1} = +\frac{1}{2}V_{Write}$, $V_{T2} = -\frac{1}{2}V_{Write}$)
2. $Z' = p$ IMP q ($V_{T1} = V_q$, $V_{T2} = V_p$)
3. Read Z'

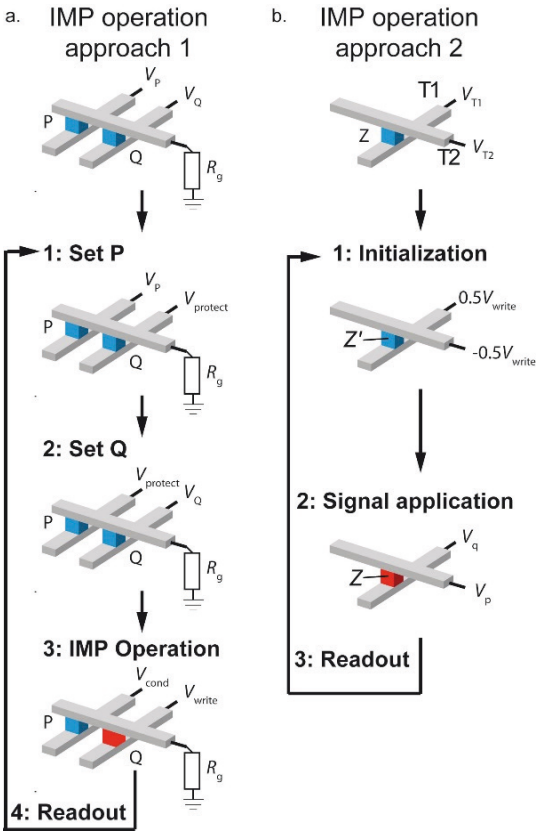


Figure 16:
(a) IMP approach according to [54] ('Stateful' logic). (b) IMP approach according to [5] (CRS logic). Blue cube represents state '0' and the red cube state '1'. From [8].

Note that complementary resistive switches can be considered finite state machines offering two states, LRS/HRS ('0') and HRS/LRS ('1') (Figure 17). By applying input signals to the two terminals T1 and T2, a switch over from '0' to '1', or vice versa can be conducted. For example, if the device is in state '0' (blue cube) the device will switch to '1' if T1 = 1 and T2 = 0 [57]. In principle 14 out of 16 Boolean functions are feasible with a single CRS cell. Note, for XOR and XNOR two cells are required.

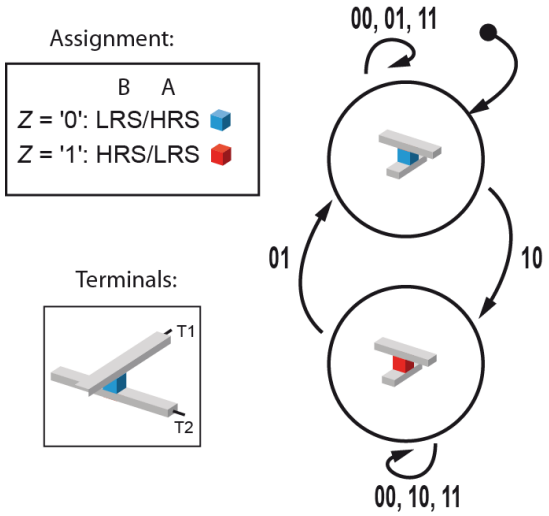


Figure 17: *Logic state and terminal assignment. Finite state machine representation of a CRS cell. From [7].*

3.2 In array computation

Passive crossbar arrays for CRS-logic

Ultra-dense ReRAM-based memory architectures will be hybrid architectures with a standard CMOS component which is responsible for controlling the passive crossbar arrays. These arrays will be fabricated on top of the CMOS layers in the backend of line (BEOL) [58]. In general, the size of the crossbar arrays should be sufficiently large to justify the control circuit overhead. Thus, either appropriate selector devices are required at each cross point, or complementary resistive switches should be applied [25].

The basic idea underlying our approach is to extend the application of hybrid CMOS/crossbar architectures from pure memory operations towards array-compatible logic-in-memory operations, by enabling a sequential access to the crossbar array devices [5]. Figure 18a depicts a possible layout. The system could consist of many arrays and one control unit, which coordinates and addresses the signals to the specific wordlines (wl) and bitlines (bl). A typical array size could be for example 128 by 128 lines. Figure 18b shows a system using CRS crossbar devices with only two arrays (A_0 and A_1) and an array size 3 by 5 to illustrate the basic concept. The structure of array A0 is depicted here. The corresponding CRS cells will be referred to as A*z*CRSwl*x*bl*y* (cmp. Figure 18b), where A*z* denotes the name of the array, in which the cell can be found, wl*x* denotes the wordline of the cell and bl*y* denotes the bitline. Thus the CRS cell A0CRSwl2bl0 is found in array A_0 at intersection wl_2 and bl_0.

Remember, CRS cells consist of two anti-serially connected ReRAM cells. (A basic CRS operation in sweep mode is depicted in Figure 19a for example.) Both logic values ‘0’ and ‘1’ are represented by an in total high resistive state, since one cell is in HRS. ’0’ is represented by LRS/HRS and ‘1’ by HRS/LRS. The ‘ON’ state is only a transition state, which is reached while changing the inner state from ‘0’to ‘1’ or back. Here a half select scheme (e.g. [59]) is applied, so that there are three different voltage levels available at the word- and bitlines, low, high and ground. The devices need steep switching kinetics, since the devices must enable switching with the maximum voltage across the device for a given time period. Additionally,

the cells must prevent switching if half of the maximum voltage is applied during the same time period. Note that a very steep switching kinetic is an intrinsic feature of resistive switching devices [21, 60], thus passive crossbar arrays are feasible.

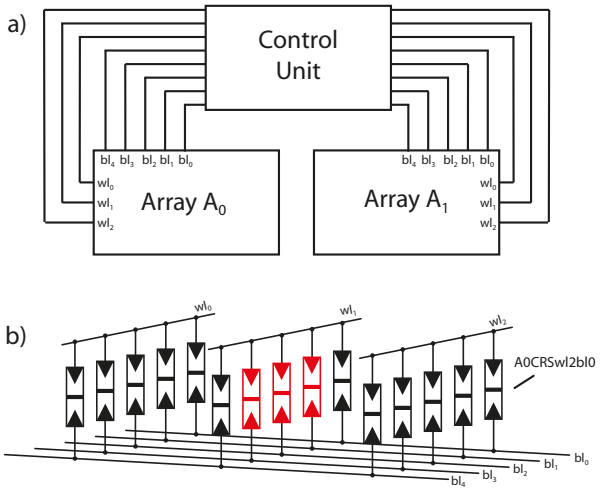


Figure 18: *Expected system section layout, which consists of two Arrays (A_0 and A_1) and a control unit. Each array has three wordlines (wl_0,wl_1 and wl_2) and five bitlines (bl_0, bl_1, bl_2, bl_3 and bl_4). The three red marked cells are used to compute a two bit addition. The control unit enables free communication between all lines and is a key element for consecutive logic. From [9].*

To switch from '0' to '1' the high potential, which is represented by the logical one '1', needs to be applied at the wordline and the low potential, logical zero '0', at the bitline of the cell. Otherwise the machine will stay in the '0'-state. To switch from '1' to '0' the low potential needs to be applied at the wordline and the high potential at the bitline of the cell. Otherwise the cell will stay in the '1'-state.

The general logic equation to represent this behavior is given by [5]:

$$Z = (wl \text{ RIMP } bl)\,Z' + (wl \text{ NIMP } bl)\,\overline{Z'} \tag{25}$$

where *wl* is the wordline connected to the device and *bl* the bitline, *Z'* is the device state prior to the application of the signals at wl and bl, and *Z* is the device state after applying the signals. As follows, if the device is in state '1' (*Z'* = '1'), the cell performs a reverse implication (RIMP) if the cell is in state '0' (*Z'* = '0') an inverse implication (NIMP) is performed. 14 out of 16 Boolean functions are directly feasible within this approach [5]. The XOR and XNOR functions can only be realized with a second CRS cell. Note that a computation on more than one device is feasible, if the wl or bl input is the same for these computations on different devices.

Equation (25) must be considered as the basic equation to develop a synthesis tool for CRS-logic.

CRS carry bit and sum bit calculation

An adder is the first step from basic logic operations towards complex arithmetic operations, since in CMOS all basic arithmetic operations (multiplier, divider and substractor) are in need of an adder. An adder consists of the possibility to calculate sum and carry bits. Figure 19c depicts the truth tables of the carry and the sum function. In these functions the actual State *Z'* is interpreted as the carry of significance i c_i, while the input variables a_i and b_i are the bits of the input words a and b with significance i. To compute c_{i+1} a_i and the negate of b_i are applied

to the wordline *wl* and bitline *bl*, respectively. Thus, using equation (25), the carry of the next higher significance c_{i+1} can be calculated by the following equation in just one step:

$$c_{i+1} = \left(a_i \text{ RIMP } \overline{b_i}\right) c_i + \left(a_i \text{ NIMP } \overline{b_i}\right) \overline{c_i} \tag{26}$$

In the next few lines it is shown that this equation offers the correct result for c_{i+1}, which is in general expressed by:

$$c_{i+1} = a_i b_i + a_i c_i + b_i c_i \tag{27}$$

This can be rewritten as follows:

$$\begin{aligned} c_{i+1} &= a_i b_i \left(c_i + \overline{c_i}\right) + a_i \left(b_i + \overline{b_i}\right) c_i + \left(a_i + \overline{a_i}\right) b_i c_i \\ &= \left(a_i + b_i\right) c_i + \left(a_i b_i\right) \overline{c_i} \\ &= \left(a_i \text{ RIMP } \overline{b_i}\right) c_i + \left(a_i \text{ NIMP } \overline{b_i}\right) \overline{c_i} \end{aligned} \tag{28}$$

Thus, the carry calculation is an intrinsic feature of the CRS-logic.

In contrast, the sum needs two steps. First, actual state *Z'* is interpreted again as the carry of significance i c_i. The input variables a_i and b_i are applied to the wordline *wl* and bitline *bl*, respectively, to calculate the intermediate state s'_i:

$$s'_i = \left(a_i \text{ RIMP } b_i\right) c_i + \left(a_i \text{ NIMP } b_i\right) \overline{c_i} \tag{29}$$

Next, c_{i+1} is required as an input signal at the bitline, while b_i is applied to the wordline:

$$s_i = \left(b_i \text{ RIMP } c_{i+1}\right) s'_i + \left(b_i \text{ NIMP } c_{i+1}\right) \overline{s'_i} \tag{30}$$

Note: It is favourable that the first sum computation step and the carry calculation step need the same input signal at the wordline, so both steps can be calculated at the same cycle in two different devices. Since the sum function needs c_{i+1} as an input signal and only a destructive read-out is available, c_{i+1} needs to be calculated in a different cell or needs to be written back.

The read-out scheme is depicted in Figure 19d. A read-out is performed by applying '1' at the wl and '0' at the bl. Due to the fact that the state can be switched from '0' to '1' (destructive readout) it is possible that a write back step is needed. If a current spike is detected in the read-out cycle, the stored information is interpreted as a '0', if no current spike occurs the information is a '1'.

Adder scheme

In this section, a way to perform multi-bit operations is introduced. Since CRS cells are passive devices there is no way, that they can pass information to the next stage. This is a major issue for complex calculations, which need more than one step or more than two input signals, like an adder. Hence either every intermediate step needs to be read out or the stored information is interpreted as a kind of 'third input' in the next step. As previously explained a read-out is destructive and requires a write back, if the data is needed later on. So the second possibility is preferable as it should be faster and more energy efficient. In fact, using parallel computing and stored information as a kind of 'third input' are the keys to designing a CRS adder.

A difficulty in realizing an adder in CRS arrays was that there is no direct XOR-functionality available in CRS-logic [5]. But as shown before (cmp. Figure 19c), it can be implemented in two steps by providing additional information from an auxiliary calculation, which is read out and used as an input signal.

More details and exemplary adder implementation can be found in [9, 61-62].

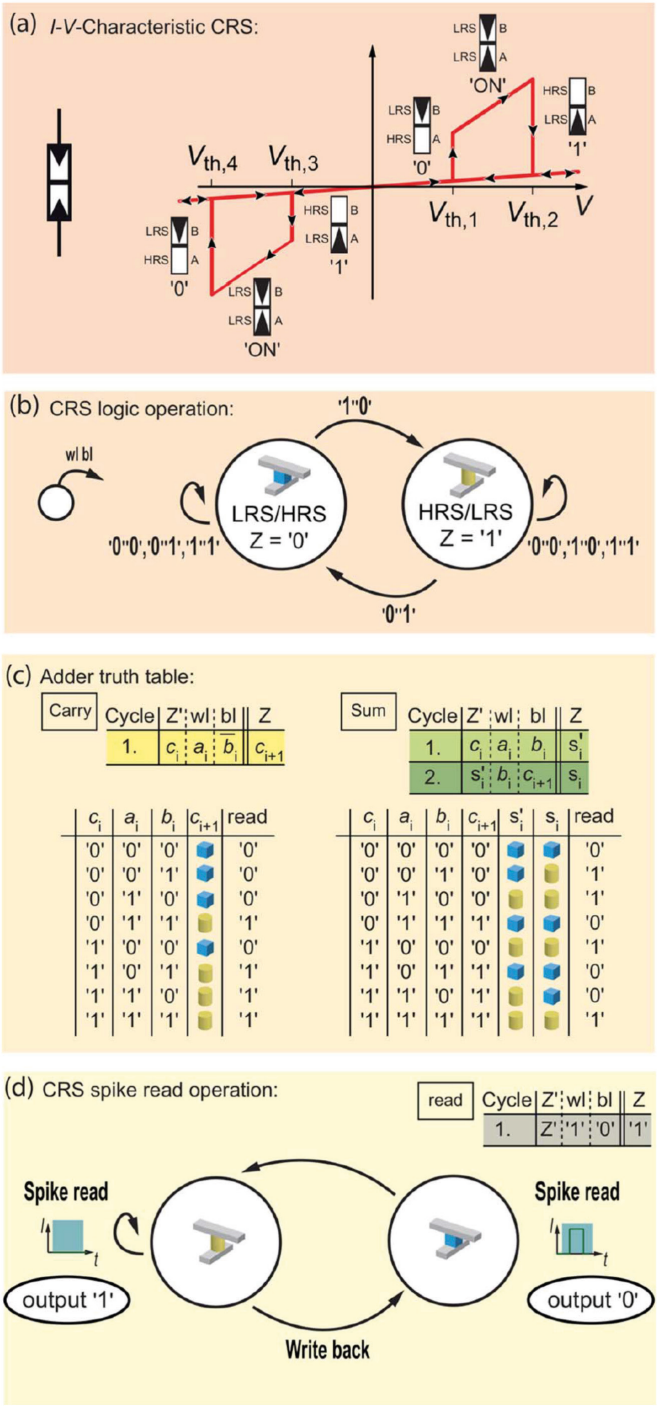


Figure 19;
(a) Basic CRS I-V-Characteristic. The logical state '0' is represented by the LRS/HRS state, logical '1' is represented by HRS/LRS and LRS/LRS is named 'ON-state' which is a transition state. The 'ON-window' is defined by $V_{th,2}$-$V_{th,1}$.
(b) CRS as a finite state machine. The inputs at wordline wl and bitline bl are a high potential, represented by a logical one '1' and low potential represented by a logical zero '0'.
(c) Truth tables for a carry and a sum functionality. The carry operation needs just one cycle (yellow), for which the actual state is interpreted as c_i and the resulting state is c_{i+1}. The sum operation needs two cycles. In the first cycle (light green) the actual state is taken as c_i and the resulting stet is interpreted as the intermediate state s'_i In the second step (dark green) the actual state is the previously calculated s'_i and the resulting state is the sum bit s_i. Note that for the second step c_{i+1} is needed as an input signal at the bitline, so c_{i+1} needs to be calculated in another cell in a previous or in the same cycle.
(d) Read-out operation (grey) for a CRS cell. A '0' was stored if a current spike (turquoise) is detected, if not it was a '1' (turquoise).

4 ReRAM-based Neuromorphic Circuits

4.1 Artificial Neural Networks

Artificial neural networks are intended to solve complex machine learning tasks by using massive parallel data processing in a similar way as in biological neural systems. A recent approach to mimic biology is to emulate the basic processing elements directly in hardware.

Animal and human nervous systems can be considered as the most successful physical realizations for processing of information. They are capable to learn, to adapt to the unexpected, to self-organize, and finally are by far the most energy-efficient computing systems by means of exhaustively exploiting parallelism. In contrast to conventional computing paradigms known from artificial digital computer architectures, the brain appears to be a non-digital, non-deterministic dynamic system comprising noisy computing elements. In this view, the result of a neural computing operation is due less to external programming and more to interacting computing elements under the strong influence of input signals. In the past decades the modelling of the brain in regard to its function has been made considerable progress, in the first instance by coming up with models for associative memories and layered neural networks. These models, however, were disregarding the property of almost any neuron, namely to emit discrete spikes for communication. Instead, *analogue* input signals were transformed into an *analogue* output signal representing an average firing rate by means of a sigmoid activation function. The emphasis in this neural network models was put on the generation of structures (i.e. "networks") by a *learning* process [63]. In particular, the learning process was organized in such a way, that the actual output of a net was brought as close as possible to a given (desired) output by minimizing an error signal. The substantial prerequisite for this optimization process was to introduce synaptic plasticity, i.e. the opportunity to modify synaptic coupling strengths using a learning rule which is driven by locally available signals such as presynaptic as well as postsynaptic actions potentials. Here, especially those synapses are strengthened whose transmitted action potentials are most successful in predicting the signal of the receiving cell. Conversely, those connections are weakened whose transmitted action potential disturb the signal of the receiving neuron considerably. However, there are very restrictive limits for a learning tasks organized in such a way. Efficient learning of a net can only be guaranteed if the presented learning patterns on one hand comprise a content of information of few 100 bits only, and on the other hand belong to the same context [64]. By increasing the size of the input patterns as well as the network size (i.e. number of neurons and synapses) a considerably increased learning effort is required to separate the significant connections from the insignificant ones which is hard to be carried out by statistical means only. The more fundamental reason for this difficulty relies in the fact that *spikeless* neural networks are unable to express *binding* between neurons. A way out of this dead end is to get the models for neurons and synapses closer to biological reality. Instead of average spike rates, spikes or whole sequences of spikes could be considered including the models for neurons and synapses which are capable to produce and to process spikes.

By consideration of a temporal signal structure it is possible to particularly take *signal correlations* into account in order to encode information. Temporal binding (in the specific form of synchronously spiking neurons) is experimentally well documented [65, 66].

The majority of neurons possess a structure which can be decomposed into three parts. Out of the *soma* sprout out a strongly branched extension the so-called *dendrite.* The dendrite serves as a physiological structure for collecting signals from other neurons and transmitting them to the receiving soma. The third part is a longer and less strongly branched extension of the soma,

the so called *axon*. Signals produced by the soma are transmitted via the axon to other neurons using their dendrites to receive these. In particular, electro-physiological measurements of in vivo neurons revealed that electrical pulses propagate along the axon featuring amplitudes of about 100mV and a pulse width of less than 2ms. Such spikes (or pulses) are termed *action potentials* [67]. Both, the onset of a spike as well as the specific shape of a spike are explained by an exchange of ions (Na^+,K^+,Ca^{2+},Mg^{2+}) through the neuron's cell membrane using dedicated ion channels. The most famous model describing the spike creation and propagation mathematically is the Hodgkin-Huxley model, which has been published in 1952 [68].

Synapses are the fundamental connection elements between neurons. Synapses propagate spikes from a presynaptic cell (a spike transmitted on a cell's axon) to a postsynaptic cell (i.e. to a cell's dendrite). If an action potential reaches a synapse so-called neurotransmitters are released into the synaptic cleft and diffuse to the synaptic spine. Here, these transmitters temporarily bond at receptor molecules. Then, the permeability of specific ion channels is momentarily influenced which results in a postsynaptic current (PSC) crossing the cell membrane of the receiving neuron. The effect of a PSC can be either excitatory or inhibitory. In case of an excitatory acting synapse an arriving action potential may cause an action potential to be generated by the receiving neuron. The PSC is then called *excitatory postsynaptic current* (EPSC). In turn, if a PSC is acting towards a suppression of an action potential the PSC is called *inhibitory postsynaptic current* (IPSC).

In particular, a synapse is modelled by an equivalent conductance where the PSC is passing through. Using $t=t_f$ as the onset of the presynaptic action potential the synaptic conductance is modelled by a time-dependent function:

$$g_S(t) = \hat{g}_S \cdot \frac{e^{-a\cdot(t-t_f)} - e^{-b\cdot(t-t_f)}}{e^{-a\ln(a/b)/(a-b)} - e^{-b\ln(a/b)/(a-b)}} \tag{31}$$

The empirical parameters a as well as b are used to fit the specific time response of a synapse. The PSC is driven by the intrinsic Nernst potential of the associated ion type and the outer membrane potential resulting in a change of the membrane potential, which is often designated as postsynaptic potential (PSP). In analogy to EPSC and IPSC the impact of a PSC to the PSP is either called *excitatory postsynaptic potential* (EPSP) if the membrane is getting depolarized or called *inhibitory postsynaptic potential* (IPSP) if the membrane potential is driven towards its resting potential. It has to be noted that postsynaptic reaction due to an incoming action potential is not always observed. In fact, the release of neurotransmitter is a statistical process [69]. In order to account for this process, a probability factor p is introduced which models the relative fraction of connections between pre- and postsynaptic cell that are *transmissive* on the average. By the combination of (4.2.1) and p an effective weight W (i.e. the synaptic efficiency) between pre- and postsynaptic cell can be defined.

Most synapses of a neuron are located at the branches of the dendritic tree. Typically, the dendritic tree is modelled as a distributed conduction net (e.g. a branched *wire*) characterized by series resistances and capacitances. Synapses situated at particular locations are modelled by shunting resistors connecting the dendrite to specific (but virtual) nodes comprising the ion-specific Nernst potentials. In most cases, the time-variant properties of the dendritic tree are, however, neglected. The synaptic inputs are considered to be active directly at the neuron's soma which leads to the so-called *point neuron model*.

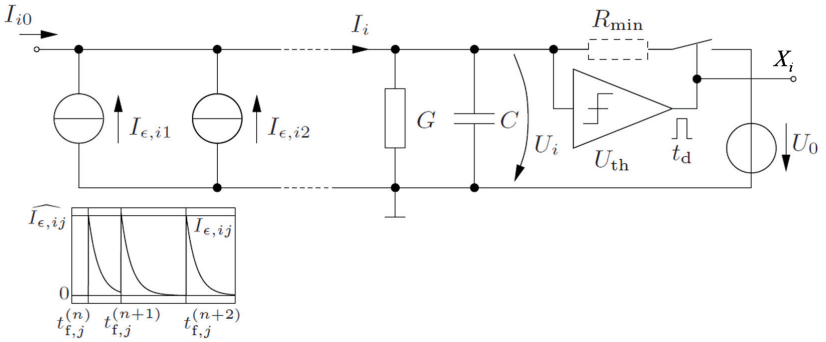


Figure 20: *Point Model of a Neuron.*

Figure 20 shows an example of a point model. The membrane of the neuron is modelled by a capacitance C which is getting charged by a total synaptic current I_i:

$$C \cdot \frac{dU_i}{dt} = G \cdot U_i + I_i \tag{32}$$

In eq. (32) U_i describes the membrane potential while G represents an ion-unspecific leakage conductance. Note, that the resting potential of the neuron has been - without restriction to the generality - arbitrarily set to $U_{\text{rest}} = 0\ V$. The total synaptic current I_i appears to be composed of individual currents generated from synapses

$$I_i(t) = I_{i0} + \sum_{j \in N_i} \sum_{n} I_{\epsilon,ij}\left(t - t_{f,j}^{(n)}\right) \tag{33}$$

In eq. (33) t, N_i, $t_{f,j}^{(n)} I_{i0}$ describe time, set of neurons presynaptically connected to neuron i, n-th firing onset of neuron j and an unspecified *analogue* input current. With respect to eq. (31) the relation

$$I_{\epsilon,ij}(t) \approx \hat{I}_S \cdot W_{ij} \cdot e^{-a\left(t - t_{f,j}\right)} \cdot \chi(t - t_{f,j}) \tag{34}$$

is often used. By integration of eq. (32) the time course of the membrane potential U_i is given. It has to be said, that eq. (32) describes a so-called subthreshold dynamics. The kinetics for generation and delivery of a spike is not included in eq. (32). By analysis of the Hodgkin-Huxley equations it turns out that eq. (32) is valid until U_i crosses a certain neuron-specific threshold U_{th}. Then, the course of the membrane potential follows a comparatively strict characteristic which is almost independent from the actual input. This is the action potential. Even though, in most artificial systems the exact pulse characteristic (i.e. the pulse shape) is not reproduced. Instead, a more or less simple pattern is used: a pulse of rectangular shape with fixed amplitude and fixed pulse duration t_d. By specifying U_{th} the onset of a pulse is given by

$$U_i(t_f) = U_{th} \quad , \quad \frac{dU_i(t_f)}{dt} > 0 \tag{35}$$

The detection of the condition eq. (35) is realized by a separate amplifier (cf. Figure 20) which generates the output pulse and initiates the discharge of the membrane capacitance C via R_{min}.

The ability of neural networks to process information obviously lies in the given connection structure, the individual effective connection strengths W_{ij} and the network's ability to establish new connections as well as to die out unnecessary connections between neurons. Especially in networks where pulses (spikes) are used to transmit information, *signal correlations* can be

taken into account for the decision which neurons should be connected and which should not. Rules for the modification of the synaptic strength which take the temporal structure of pulse activity into account are summarized under the term *Spike-Timing Dependent Plasticity* (STDP). STDP can appear on various time scales. Rules causing particular connections to be established persistently (e.g. long term potentiation, LTP [70]) are associated with a learning process. On the other hand synapses exist that show an extremely short-term persistence in its synaptic efficiency [71, 72].

Synaptic plasticity is reflected by a change of the synaptic efficiency W_{ij}. Typically, the change is mathematically described by a differential equation. Any rule describing the change of W_{ij} takes only *local* signals and states into account in order to keep biological plausibility. For synaptic plasticity based on STDP the presynaptic action potential X_j, postsynaptic action potential X_i, the postsynaptic membrane potential U_i, and derived quantities can be considered as driving forces for changing W_{ij}, cf. Figure 21.

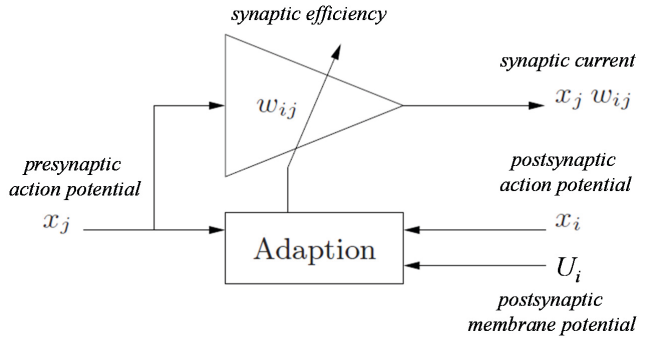


Figure 21: *Change of synaptic efficiency as a function of local signals, from [73].*

A simple STDP rule fulfilling Hebb's postulate [74] is given by

$$\tau \cdot \frac{dW_{ij}}{dt} = -W_{ij} + \mu \cdot (U_i - B) \cdot \chi(X_j) \qquad W_{ij} \overset{!}{\geq} 0\ .\ [73] \tag{36}$$

In eq. (36) τ, B, μ, X_j, U_i describe a time constant, an adaptation threshold ($0 < B < U_{th}$), an adaptation factor, the presynaptic action potential, and the postsynaptic membrane potential. If τ is small compared to the average spike interval of X_j the weight W_{ij} initially decays to almost 0. If the membrane potential U_i is close to the receiving neuron's threshold (e.g. by receiving input from a different functional layer) and the presynaptic neuron fires, the weight W_{ij} rises exponentially (note that a feedback path exists between W_{ij} and U_i by eq. (32)) causing the receiving neuron to fire *synchronously* to X_j. In the case that U_i is significantly lower than B, the weight W_{ij} decays rapidly to 0 which results in a completely decoupled state. In other words, synapses of type eq. (36) tend to *synchronize* neurons, in case that they encode similar information [73]. Synapses of type eq. (36) were successfully used to implement various functional layers of early processing stages of an artificial vision system. Emphasis is to be laid on the implementation of orientation-sensitive features detectors based on Gabor-wavelets, segmentation of grey-level encoded images, and feature binding [73, 75-77].

By enlarging τ (i.e. enlarging of the time scale) the subthreshold characteristics represented by U_i can be replaced by the action potential X_i. Then, equation (36) is transformed into

$$\tau \cdot \frac{dW_{ij}}{dt} = \mu \cdot \chi\left(X_i \cdot X_j\right) + F\left(X_i, X_j, W_{ij}, \ldots\right) \tag{37}$$

In eq. (37) the co-existence of spikes results in a strengthened coupling of neuron *i* and neuron *j* in a long-term view, which was the original intention described by Hebb [74]. Modifications of eq. (37) were successfully applied for various applications including associative memories [78], hebbian based maximum eigenfilter (Oja's rule and principal component analysis), independent component analysis [63], and much more besides these.

4.2 CMOS neuromorphic circuits

So far, the computational complexity of realistic neuron and synapse models prevented the simulation of large scale networks on standard computer systems making it necessary to come up with dedicated hardware architectures, which are capable to simulate functional networks in real time. In the past decades major effort has been made to mimic the behaviour of biological neurons and synapses by means of integrated circuits based on CMOS technology.

In the following a set of fundamental CMOS circuits is shown which are frequently used in neuromorphic circuits. The main purpose is to illustrate the requirements for the implementation of spiking neurons and hebbian synapses. Though this overview must be incomplete, last but not least due to the exhaustive literature that has been published over the years, it clearly demonstrates the persistence of a gap between circuit complexity induced by synaptic dynamics and CMOS scalability. In order to close this gap, new device concepts need to be developed. Especially ReRAM devices are considered as key elements to realize highly scalable and low-power neuromorphic systems consist using a hybrid analog-digital circuit approach. The specific properties of ReRAM devices that make those devices highly useful as artificial synapse are highlighted in detail. Especially, the multi-level capability of ReRAM devices enables the implementation of learning rules such as Spike-Timing Dependent Plasticity (STDP). The scaling perspectives of ReRAM based neuromorphic architectures are elaborated on, revealing a scaling potential below 10 nm.

The simplest circuit for the implementation of a neuron model capable of receiving and generating pulses is shown in Figure 22. The circuit consists of a threshold switch and a capacitor formed by the gate-bulk capacitance of a transistor. The gate-bulk capacitor of a MOS transistor is an optimal realization of the membrane capacitor with regard to area.

In order to meet power figures comparable with biological neurons typical currents in a circuit implementation are bound to several $x10^{-8}$A. The threshold switch should have a large hysteresis in order to exploit the voltage range defined by the supply voltage to its full extent and minimize the area of the membrane capacitor. In concurrent CMOS technology the voltage difference between the resting potential after firing and the switching threshold is limited to values of about 0.5V up to 1.0V. Applying a charge current of 10nA a capacitor of 100pF would be required to model the dynamic properties of a cell membrane under the assumption of an integration time of 10ms (equivalent to a pulse frequency of 100 Hz). This capacitor becomes very inefficient in regard to area occupation (i.e. in 130nm CMOS technology an array of 16.000 neurons would have an area occupation of $320mm^2$). It is therefore necessary to scale up the typical (expected) firing frequency and to scale down the pulse duration t_d in order to obtain reasonable figures for the required area. In order to limit the active power consumption of the circuit, the average pulse frequency has to be bound to an upper limit. Reasonable values

for the highest pulse frequency are between 1 MHz and 10 kHz which keeps the order of magnitude for power consumption compared to biological neurons. It follows that typical implementations of pulsing neurons rely on a membrane capacitance considerably less than 1pF but still larger than 100 fF [73][77a].

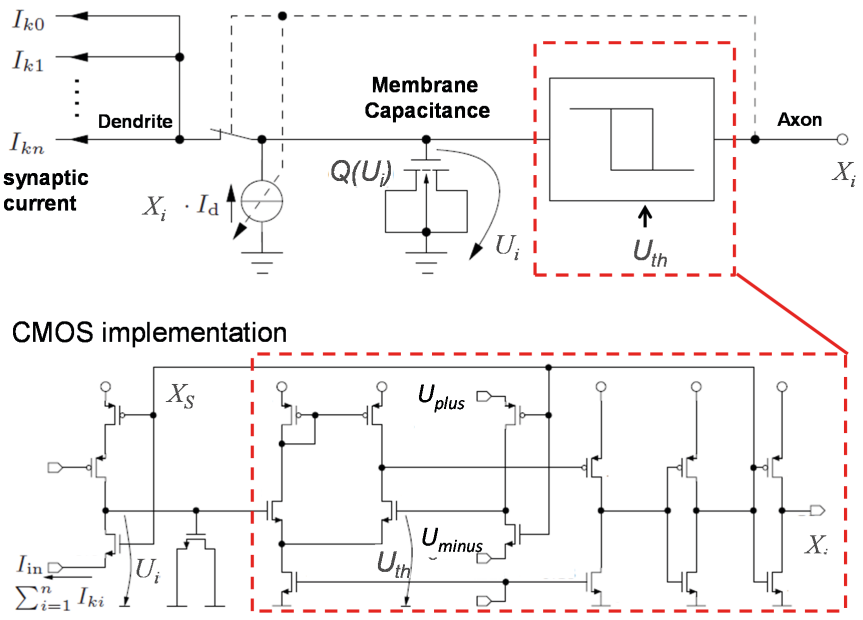


Figure 22: *Block Diagram and CMOS circuit implementation of a Neuron, from [73].*

A threshold switch featuring positive feedback is shown in Figure 22, within the dashed area. The first stage is composed of a differential pair with current mirror. The tail current of the differential pair flows continuously. Due to power consumption this current must be chosen as small as possible. Conversely, by reduction of the tail current the switching speed of the circuit gets lost. Especially a delay between the input signal crossing the threshold and the rising edge of the output signal can be expected at low tail currents causing the circuit to operate in an unstable regime. An acceptable compromise can be found for a tail current of 20nA (in this example for a 130nm CMOS technology).

The following stage is a PMOS source stage with constant source current load. This stage significantly contributes to the voltage gain of the amplifier. The last stage is realized with an inverter comprising minimal dimensions. Its input signal is amplified enough in order to ensure, that the voltage range causing a cross current through the inverter is passed sufficiently fast.

Finally, the positive feedback is realized with a bipolar switch either the upper (U_{plus}) or lower (U_{minus}) threshold voltage is passed to the non-inverting input of the amplifier. In the receiving phase of the neuron U_{minus} is continuously compared with the membrane voltage.

By reaching the threshold the state of the output signal flips and the membrane is charged to the level of U_{plus}, cf. Figure 23.

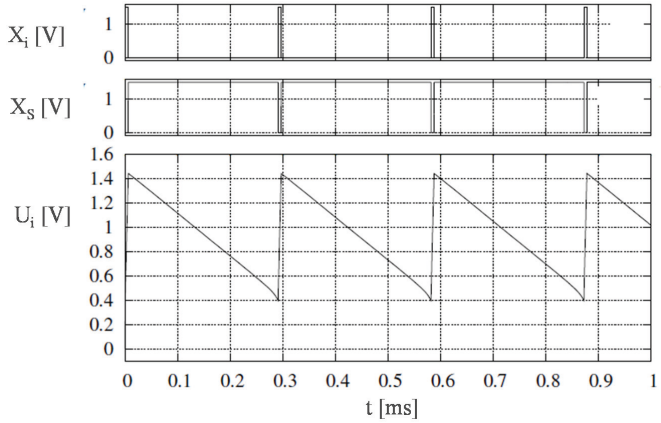


Figure 23: *Pulse generation using a constant PSC, from [73].*

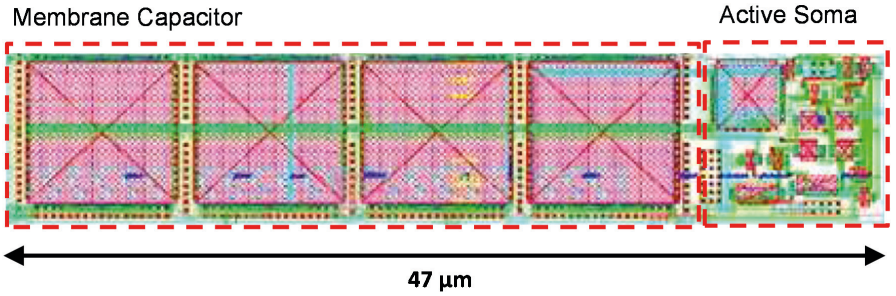


Figure 24: *Layout of an artificial neuron in 130 nm CMOS technology. Roughly 80% of the layout area is occupied by the membrane capacitance.*

Figure 24 shows the layout of a neuron circuit. It clearly shows that the membrane capacitor represents the main area contribution of the circuit. Because the number of neurons in a neural net is moderate (compared to the number of required synapses) this area effort is acceptable.

Figure 25 shows the fundamental circuit of a dynamic synapse. The weight W_{ij} is represented by a voltage U_{weight} which is caused by charges stored on C_{weight}. Dependent on U_{weight} transistor M_{N0} delivers a current proportional to its gate-source voltage which requires M_{N0} to be operating in the triode regime. An incoming pulse X_j turns on transistors M_{N3} and M_{N4}. The circuit composed of M_{N5} and M_{N6} realize a current divider. I.e. the synaptic output current I_{ji} is a defined fraction of the drain current delivered by M_{N0}. The current ratio depends on the externally applied voltages U_{cdiv1} and U_{cdiv2}. The reason for choosing a current divider is given by the fact that M_{N0} cannot deliver currents below a limit defined by the triode regime of operation which is approximately found at 100nA. Smaller synaptic currents have to be derived from the drain current I_D by dividing I_D using a fixed ratio. Here, synaptic output currents less than 2nA can be realized in a stable way.

Note that M_{N5} and M_{N6} represent a differential pair whose tail current is modified by changing U_{weight}. This change is achieved by delivering a current I_{adapt} to C_{weight}. Synaptic plasticity can be realized by introduction of a specific circuit delivering a particular current in relation to the chosen adaptation (or learning) rule. Figure 26 shows a circuit implementing an approximate behaviour with regard to the rule eq. (36).

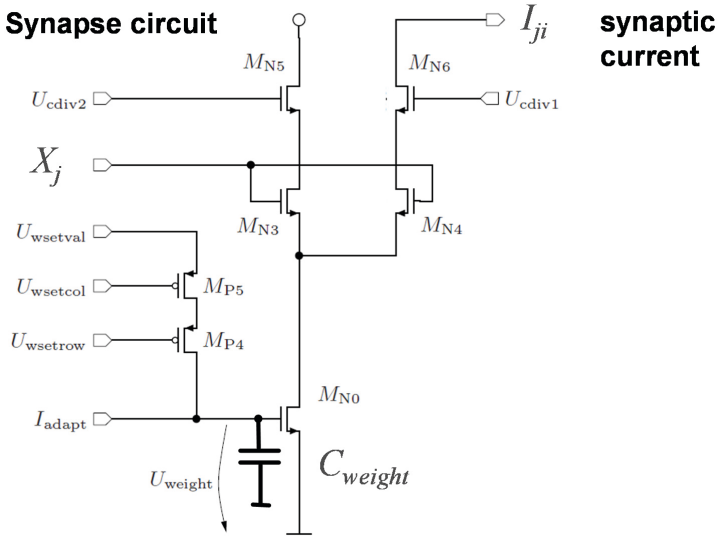


Figure 25: *CMOS circuit of a dynamic synapse.*

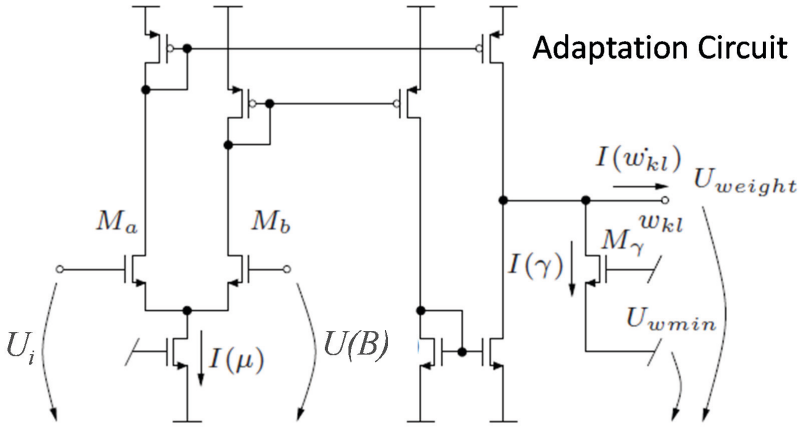


Figure 26: *CMOS implementation of a specific learning rule, from [73].*

Basically, transistor M_a and M_b form a differential pair. While the input voltage for transistor M_b represents the parameter B in eq. (36) the input voltage of transistor M_a represents the membrane potential of the receiving neuron. Using current mirrors the output current of the circuit is the difference of the drain currents delivered from M_a and M_b. For small voltage differences $(U_i - U(B))$ the output current is a linear function of the voltage difference while for larger voltage differences the output current saturates to the tail current $I(\mu)$ (keeping the sign of the voltage difference) which specifies the maximum adaptation rate. The transistor $M\gamma$ represents the leakage term in eq. (36). By keeping the requirements for signal currents comparable to the specifications made for the neuron circuit, C_{weight} is found in an order of magnitude similar to the membrane capacitance. However, if the weight capacitor is reduced in its capacity a faster – and hence a more bipolar-oriented – characteristics is obtained. Learning, in contrast to adaptation, is a process which operates on a larger time scale and has a need for persistent representation of memory states. It has to be concluded, that a small-sized persistent (non-volatile) memory element needs to be found. Since the number of synapses in a neuromorphic system is by roughly 3 orders of magnitude larger than the number of neurons the synapse circuit should be as simple as possible. It should comprise an inherent dynamics for weight change, it should comprise multilevel capabilities, and weight storage should be non-volatile unless a change in the synaptic efficiency is necessary in order to improve the system performance (cmp. Figure 27).

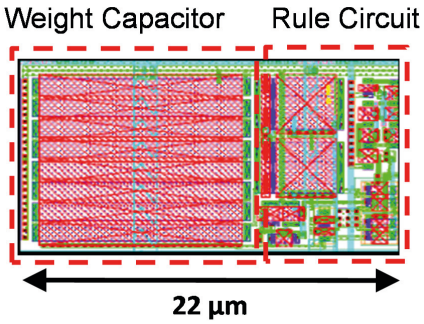


Figure 27: *Layout of an adaptive Synapse in 130nm CMOS technology.*

4.3 ReRAMs for neuromorphic circuits

Most recent neuromorphic approaches are based on pure CMOS or SOI technology. However, the soon projected availability of novel nano devices [58] offers some unique properties which are highly advantageous for implementing hardware synapses [79].

Depending on each approach one or more of the following properties are exploited [2]:

- Non-volatility / Volatility of resistive states
- Non-linear switching kinetics
- Multilevel resistance behavior
- Switching statistics
- Capacitive properties

The basic idea in most ReRAM-based neuromorphic approaches is to consider ReRAM devices, or small ReRAM-based circuits, as artificial synapses. One of the first ideas to use Re-

RAM devices for neuromorphic applications goes back to Likharev [80], introducing the concept of 'Crossnets' which uses ReRAM devices as programmable interconnects, i.e. binary synapses. On top of CMOS-based neurons, a crossbar array of nano-scaled ReRAM devices is used for reconfigurable wiring of the neurons. Binary ReRAM synapses can also be directly used to implement associative memories [81, 82]. Note that distinct *I-V* non-linearity of the LRS branch is required in this approach, which is difficult to achieve with common ReRAM devices [83]. Thus, a serial selector is usually required.

Further concepts consider the tunable resistive state of ReRAM devices as synaptic weight. By doing so, synaptic plasticity rules can be implemented. In biological systems, neurons communicate between them through synapses, which are characterized by a weight, as schematized in Figure 28a. By updating the weight of the synapses, the communication between a pre-neuron (before the synapse) and a post-neuron (after the synapse) changes, enabling the possibility of implementing biological operations such as pattern learning and recognition. The weight of the synapses is updated following a learning rule called Spike Timing Dependent Plasticity, or STDP, which depends on the relative timing of the electrical pulses arriving from the pre-neuron and the post-neuron on the synapse. In case the pre-neuron spikes before the post- neuron, the synapse conductance is enhanced, while, in case the post-neuron spikes before the pre-neuron, the synapse is depressed, hence the synapse conductance is decreased. The biological results are shown in Figure 28b for the case of a rat-hippocampal neuron, showing an approximately exponential behaviour. In 2008 G. Snider suggested to implement STDP through the use of memristors as synapses since the gradual conductance modulation of the memristor makes it a promising nanoscale device for emulating synapses in artificial neural networks [84]. A single spike cannot alter the resistive state, and weight update is induced when a post and pre synaptic pulse overlap. Note that the pulse width of the spikes is decreased successively with time according to an exponential law. Moreover, the synaptic weight is either increased or decreased, depending on whether excitatory or inhibitory inputs are applied.

Interestingly, the exemplary STDP approach of Snider uses three specific ReRAM properties, namely non-volatility of resistive states, non-linear switching kinetics and multilevel resistance behaviour, whereas the Crossnet approach only requires non-volatility.

The first experimental evidence of the feasibility of the STDP learning curve with memristors was in 2010 by S. H. Jo, et al. [85]. The conductance update is reported in Figure 28c and it shows a conductance increase corresponding to a pre-neuron that spikes before a post-neuron, hence $\Delta t = t_{\mathrm{pre}} - t_{\mathrm{post}} < 0$. Conversely, when the pre-neuron spikes after the post-neuron, the memristor conductance is decreased, corresponding to a $\Delta t < 0$. The obtained STDP curve shows an exponential behaviour which is consistent with biological data in Figure 28b.

Non-volatility and Volatility of Resistive States

The non-volatility of the resistive states, i.e. the hysteretic memristive behavior, is the basic ReRAM property which enables synaptic functionality. In general, ReRAM devices offer long-term state retention up to ten years at typical temperatures below 85 °C [86]. However, this is not true for all ReRAM devices and all the possible resistive states. In [87] the feasibility of short-term plasticity (STP) and long-term potentiation (LTP) was suggested for Ag_2S based devices. The volatility of high ohmic ON states enables implementation of STP within a single ReRAM device depending on the spike rate, whereas the non-volatility of the permanent LRS enables implementation of LTP. Similarly, internal non-equilibrium states which cause an emf (electromotive force) voltage [88] may also influence the resistance states [89]. This means that ReRAM states can offer exponential forgetting, i.e., the resistive state is varied exponentially with time, depending on internal non-equilibrium states, for example.

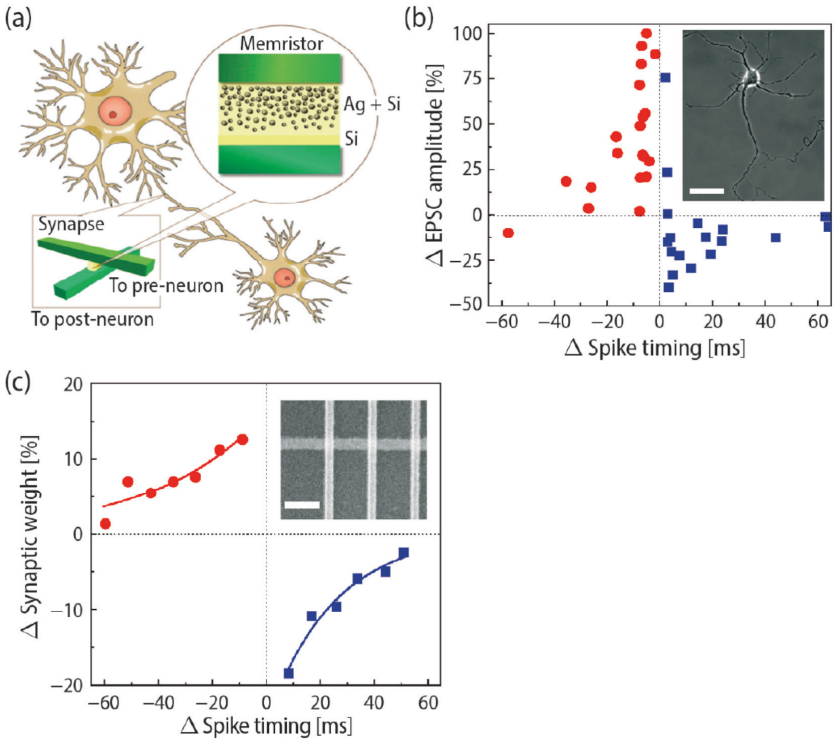


Figure 28: *(a) Schematic description of the role of the memristor as a synapse between two neurons. (b) shows the experimental change of excitatory postsynaptic current (EPSC) of rat hippocampal neurons as a function of the relative spike timing. (c) Experimental memristor STDP curve. The exponential behaviour is similar to biological measurements. From [85].*

Non-linear switching kinetics

ReRAM devices in general offer a highly non-linear exponential switching kinetics [21, 90], i.e., the set time t_{SET} exponentially depends on the applied pulse height. This feature can be used to implement Spike-Timing Dependent Plasticity (STDP) in a simple manner [84, 91].

For a certain pulse length, one can define a threshold voltage below which no switching occurs for the specific pulse duration. If we select a voltage V_{pulse} as pre-synaptic spike voltage and -V_{pulse} as post synaptic spike voltage, which are both well below the threshold voltage, the application of either the pre-synaptic or post-synaptic pulse alone would lead to no or only a slow synaptic weight adaption. If pre or post synaptic signals occur simultaneously instead, the total voltage at the junction will be equal to $2V_{pulse}$ which shall be well above the threshold voltage. In this case, therefore, the synaptic weight adaption will be fast, thus enabling STDP. Figure 29 shows a corresponding STDP implementation according to [91].

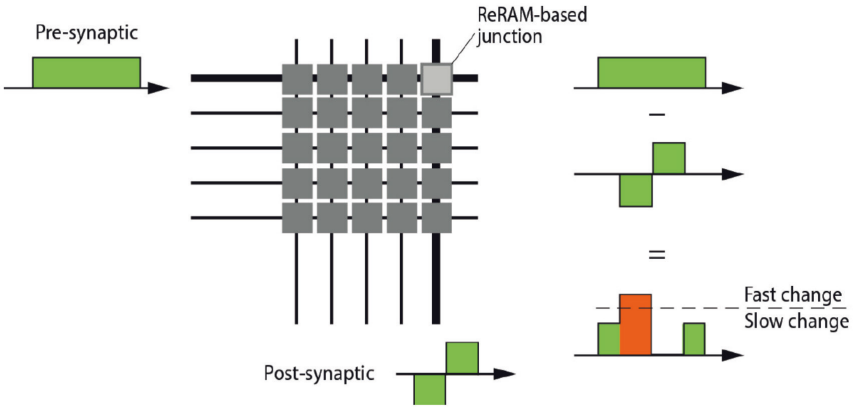


Figure 29: *(a) Implementation of simple STDP functionality using a ReRAM-based cross-bar array. In this straightforward approach two types of pre- and post-synaptic signals, both being below the threshold voltage, are considered. Only if the post-synaptic signal occurs while the pre-synaptic signal is still active, the sum signal is large enough to induce a fast change of the resistive state. If either the pre-synaptic or the post-synaptic signal is active, no or only a slow change of resistance state occurs. This approach is described in [91], for example. From [85], chapter 25.*

In [92] a similar STDP approach is suggested, but the spike's shape is more accurate in terms of biological signals. A recent review on STDP using memristive devices is [93] where also the impact of ReRAM device models is discussed.

Multilevel resistance behavior

Another property of ReRAM devices is the capability of multi-level resistances [94-97]. Multiple levels offer the possibility to have multi-bit synapses instead of binary synapses.

Moreover, there are also reports on material systems which show a gradual resistance change, hence offer a gradual adaption of a synaptic weight [85]. Gradual tunable ReRAM devices offer the potential to mimicking biological synaptic behaviour more realistically than bistable devices (e.g. SRAM-based synapses). However, for typical ReRAM devices, the SET process is an abrupt process, and either a current compliance or a series resistor is required to obtain a certain multi-level or gradual tunable state, respectively. During SET a current compliance (e.g. provided by a transistor) or control of the RESET voltage V_{stop}.

Switching statistics

Another highly interesting property of ReRAM devices is connected to the device's switching statistics [79]. This property can be used to emulate the non-deterministic synaptic weight update also known from biological synapses. Note that switching statistics of ReRAM devices is linked with the non-linear switching kinetics of the ReRAM device: Depending on the applied signal amplitude and pulse length, the mean SET/RESET operation time is varied. Moreover, the variance also depends on the voltage operation regime, i.e., the switching process which determines the kinetics.

This feature might be useful to implement learning tasks, e.g., probabilistic STDP [98]. There an Ag-based ECM cell in 1T1R (one transistor + 1 resistive switch) configuration is considered. By applying so called 'weak' switching conditions, a certain switching probability, e.g. 50 %, can be realized. Especially, for ReRAM devices offering an abrupt SET process, exploiting switching statistics might be easier than using the multi-level capabilities because simpler control circuitry can be applied.

Capacitive properties

In general, ReRAM devices are considered resistive devices. However, ReRAM devices consist of a metal/insulator/metal (MIM) structure, thus also offer a capacitance in parallel [99]. This capacitance can be exploited when considering CRS cells as binary synapses (Figure 30). The resistance of element A and B is switchable between the HRS and LRS, whereas the capacitance is completely specified by its dimensions and the permittivity of the insulator material. For the logic state '0', element A is in the LRS, and thus the capacitance C_A is short-circuited. In this case the overall capacitance is determined by C_B.

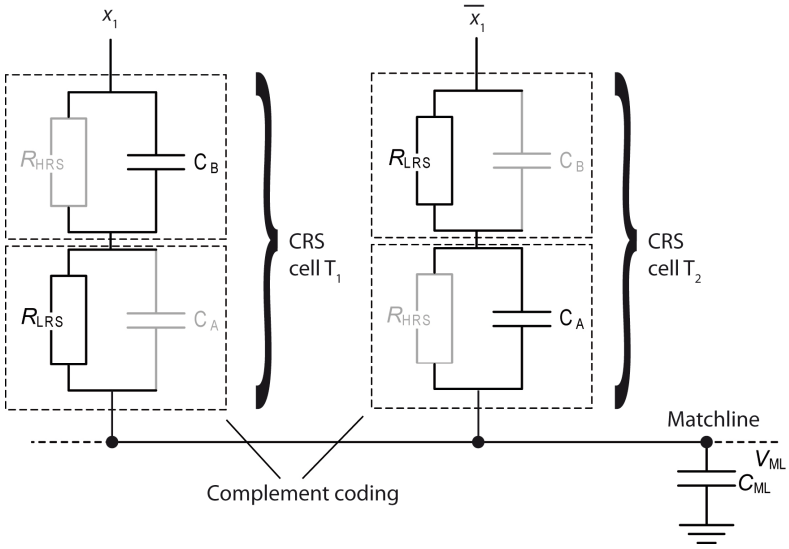


Figure 30: *Equivalent circuit model of an associative capacitive network based on CRS cells. Each CRS cell consists of two part cells A and B, offering different capacitances C_A and C_B. 2M CRS cells are connected to each matchline. There are N matchlines where the Hamming Distance (corresponds to V_{ML}) between input and stored templates is evaluated in parallel. From [7].*

Similarly, for state '1', element B is short-circuited and capacitance C_A dominates the overall device behavior. In result, the stored information of the CRS cell influences the detectable capacitance, which thus can be read out non-destructively. In Figure 30 a capacitive voltage divider is formed with C_{ML}, and V_{ML} is evaluated using a sense amplifier. The switchable capacitance feature of this device can be used to extend the functionality of the binary synapse and enables pattern recognition tasks, so called Associative Capactive Networks [100].

To enable a proper matching operation a complement coding is applied: CRS cell T_1 stores the information bit and CRS cell T_2 stores the negate (see Figure 30). Correspondingly, both the search pattern (x_1 ... x_M) and its complement are applied to the CRS array. Thus, when considering N stored patterns, the array size is N x 2M. Finally, the similarity of input patterns (x_1 … x_m) and stored patterns is evaluated in parallel on each matchline (ML). Note that in an associative capacitive network, CRS cells can be considered binary synapses while the CMOS comparator circuit offers the Neuron functionality.

References

[1] E. Linn, "Complementary Resistive Switches,", Phd thesis, 2012.

[2] D. Ielmini and R. Waser, *Resistive Switching - From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications* Wiley-VCH, 2016.

[3] E. Linn, A. Siemon, R. Waser, and S. Menzel, "Applicability of Well-Established Memristive Models for Simulations of Resistive Switching Devices," *IEEE Transactions on Circuits and Systems - Part I: Regular Papers (TCAS-I)*, vol. 61, pp. 2402 - 2410, 2014.

[4] A. Siemon, S. Menzel, A. Marchewka, Y. Nishi, R. Waser, and E. Linn, "Simulation of TaO_x-based Complementary Resistive Switches by a Physics-based Memristive Model," *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1420 - 1423, 2014.

[5] E. Linn, R. Rosezin, S. Tappertzhofen, U. Böttger, and R. Waser, "Beyond von Neumann-logic operations in passive crossbar arrays alongside memory operations," *Nanotechnology*, vol. 23, pp. 305205, 2012.

[6] L. Nielen, A. Siemon, S. Tappertzhofen, R. Waser, S. Menzel, and E. Linn, "Study of Memristive Associative Capacitive Networks for CAM Applications," *IEEE J. Emerging Sel. Top. Circuits Syst.*, vol. 5, pp. 153-161, 2015.

[7] E. Linn, "Memristive Nano-Crossbar Arrays Enabling Novel Computing Paradigms," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2596-2599, 2014.

[8] E. Linn, "Memristive Devices - The key enabler for CIM architecture implementation," *DATE Conference 2015*, 2015.

[9] A. Siemon, S. Menzel, R. Waser, and E. Linn, "A Complementary Resistive Switch-based Crossbar Array Adder," *IEEE J. Emerging Sel. Top. Circuits Syst.*, vol. 5, pp. 64 - 74, 2015.

[10] L. O. Chua, "Nonlinear circuit foundations for nanodevices, part I: The four-element torus," *Proc. IEEE*, vol. 91, pp. 1830-1859, 2003.

[11] L. Chua, "Device Modeling via Basic Non-Linear Circuit Elements," *IEEE Transactions on Circuits and Systems*, vol. 27, pp. 1014-1044, 1980.

[12] H. Marquez, *Nonlinear Control Systems: Analysis and Design* Wiley-Interscience, 2003.

[13] L.O. Chua and S.M. Kang, "Memristive devices and systems," *Proc. IEEE*, vol. 64, pp. 209-223, 1976.

[14] Y. V. Pershin and M. Di Ventra, “Memory effects in complex materials and nanoscale systems,” *Adv. Phys.*, vol. 60, pp. 145-227, 2011.

[15] L.O. Chua, “Resistance switching memories are memristors,” *Appl. Phys. A-Mater. Sci. Process.*, vol. 102, pp. 765-783, 2011.

[16] L.O. Chua, “Memristor-the missing circuit element,” *IEEE Trans. Circuit Theory*, vol. CT-18, pp. 507-519, 1971.

[17] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, “The missing memristor found,” *Nature*, vol. 453, pp. 80-83, 2008.

[18] J. Mustafa and R. Waser, “A novel reference scheme for reading passive resistive crossbar memories,” *IEEE Trans. Nanotechnol.*, vol. 5, pp. 687-691, 2006.

[19] R. E. Pino, J. W. Bohl, N. McDonald, B. Wysocki, P. Rozwood, K. A. Campbell, A. Oblea, and A. Timilsina, “Compact method for modeling and simulation of memristor devices: Ion conductor chalcogenide-based memristor devices,” *IEEE/ACM International Symposium on Nanoscale Architectures*, pp. 1-4, 2010.

[20] R. Waser, R. Dittmann, G. Staikov, and K. Szot, “Redox-Based Resistive Switching Memories - Nanoionic Mechanisms, Prospects, and Challenges,” *Adv. Mater.*, vol. 21, pp. 2632-2663, 2009.

[21] S. Menzel, M. Waters, A. Marchewka, U. Böttger, R. Dittmann, and R. Waser, “Origin of the Ultra-nonlinear Switching Kinetics in Oxide-Based Resistive Switches,” *Adv. Funct. Mater.*, vol. 21, pp. 4487-4492, 2011.

[22] Y. Nishi, S. Schmelzer, U. Böttger, and R. Waser, “Weibull Analysis of the Kinetics of Resistive Switches based on Tantalum Oxide Thin Films,” *Proceedings of the 43rd European Solid-State Device Research Conference (ESSDERC)*, pp. 174-177, 2013.

[23] S. Yu, Y. Wu, and H. Wong, “Investigating the switching dynamics and multilevel capability of bipolar metal oxide resistive switching memory,” *Appl. Phys. Lett.*, vol. 98, pp. 103514/1-3, 2011.

[24] F. Alibart, L. Gao, B. D. Hoskins, and D. B. Strukov, “High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm,” *Nanotechnology*, vol. 23, pp. 75201, 2012.

[25] E. Linn, R. Rosezin, C. Kügeler, and R. Waser, “Complementary Resistive Switches for Passive Nanocrossbar Memories,” *Nat. Mater.*, vol. 9, pp. 403-406, 2010.

[26] S. Schmelzer, E. Linn, U. Böttger, and R. Waser, “Uniform Complementary Resistive Switching in Tantalum Oxide Using Current Sweeps,” *IEEE Electron Device Lett.*, vol. 34, pp. 114-116, 2013.

[27] E. Linn, S. Menzel, R. Rosezin, U. Böttger, R. Bruchhaus, and R. Waser,*Nanoelectronic Device Applications Handbook: Modeling of Complementary Resistive Switches* CRC Press, Taylor & Francis group, 2013, pp. 315-325.

[28] S. Benderli and T. A. Wey, “On SPICE macromodelling of TiO_2 memristors,” *Electronics Letters*, vol. 45, pp. 377-379, 2009.

[29] Y. N. Joglekar and S. J. Wolf, “The elusive memristor: properties of basic electrical circuits,” *Eur. J. Phys.*, vol. 30, pp. 661-675, 2009.

[30] Z. Biolek, D. Biolek, and V. Biolkova, “SPICE Model of Memristor with Nonlinear Dopant Drift,” *Radioengineering*, vol. 18, pp. 210-214, 2009.

[31] S. Shin, K. Kim, and S. M. Kang, “Compact Models for Memristors Based on Charge-Flux Constitutive Relationships,” *IEEE Trans. Comput-Aided Des. Integr. Circuits Sys*, vol. 29, pp. 590-598, 2010.

[32] D. Biolek, Z. Biolek, and V. Biolkova, “Pinched hysteretic loops of ideal memristors, memcapacitors and meminductors must be ‘self-crossing’,” *Electronics Letters*, vol. 47, pp. 1385 - 1387, 2011.

[33] F. Corinto and A. Ascoli, “A Boundary Condition-Based Approach to the Modeling of Memristor Nanostructures,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, pp. 2713-2726, 2012.

[34] T. Prodromakis, B. P. Peh, C. Papavassiliou, and C. Toumazou, “A Versatile Memristor Model With Nonlinear Dopant Kinetics,” *IEEE Trans. Electron Devices*, vol. 58, pp. 3099-3105, 2011.

[35] E. Linn, S. Menzel, R. Rosezin, U. Böttger, R. Bruchhaus, and R. Waser, “Modeling Complementary Resistive Switches by Nonlinear Memristive Systems,” *Proceedings of the 11th IEEE Conference on Nanotechnology*, pp. 1474-1478, 2011.

[36] R. K. Budhathoki, M. P. Sah, S. P. Adhikari, H. Kim, and L. Chua, “Composite Behavior of Multiple Memristor Circuits,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, pp. 2688-2700, 2013.

[37] F. Corinto, A. Ascoli, and M. Gilli, “Analysis of current–voltage characteristics for memristive elements in pattern recognition systems,” *International Journal of Circuit Theory and Applications*, vol. 40, pp. 1277-1320, 2012.

[38] H. Kim, M. Pd. Sah, C. Yang, T. Roska, and L. O. Chua, “Memristor Bridge Synapses,” *Proceedings of the IEEE*, vol. 100, pp. 2061-2070, 2012.

[39] Y. V. Pershin and M. Di Ventra, “Solving mazes with memristors: A massively parallel approach,” *Phys. Rev. E: Stat. Nonlinear Soft Matter Phys.*, vol. 84, pp. 46703/1-6, 2011.

[40] Y. Ho, G. M. Huang, and P. Li, “Dynamical Properties and Design Analysis for Nonvolatile Memristor Memories,” *IEEE Trans. Circuits Syst. I-Regul. Pap.*, vol. 58, pp. 724-736, 2011.

[41] J. H. Hur, M.-J. Lee, C. B. Lee, Y.-B. Kim, and C.-J. Kim, “Modeling for bipolar resistive memory switching in transition-metal oxides,” *Phys. Rev. B*, vol. 82, pp. 155321-, 2010.

[42] J. R. Heath, P. J. Kuekes, G. S. Snider, and R. S. Williams, “A Defect-Tolerant Computer Architecture: Opportunities for Nanotechnology,” *Science*, vol. 280, pp. 1716-21, 1998.

[43] K. K. Likharev and D. B. Strukov, “CMOL: Devices, Circuits, and Architectures,” *Introducing Molecular Electronics*, vol. 680, pp. 447-477, 2006.

[44] A. Dehon, “Nanowire-Based Programmable Architectures,” *ACM Journal on Emerging Technologies in Computing Systems*, vol. 1, pp. 109-162, 2005.

[45] J. L. Kouloheris and A. El Gamal, “PLA-based FPGA area versus cell granularity,”, pp. 4.3/1-4,1992.

[46] M. R. Stan, P. D. Franzon, S. C. Goldstein, J. C. Lach, and M. M. Ziegler, “Molecular electronics: from devices and interconnect to circuits and architecture,” *Proc. IEEE*, vol. 91, pp. 1940-1957, 2003.

[47] G. S. Rose and M. R. Stan, “A programmable majority logic array using molecular scale electronics,” *IEEE Trans. Circuits Syst. I-Regul. Pap.*, vol. 54, pp. 2380-2390, 2007.

[48] G. S. Snider and P. J. Kuekes, “Nano state Machines using hysteretic resistors and diode crossbars,” *IEEE Transactions on Nanotechnology, USA*, vol. 5, pp. 129-137, 2006.

[49] G. S. Snider and R. S. Williams, “Nano/CMOS architectures using a field-programmable nanowire interconnect,” *Nanotechnology*, vol. 18, pp. 035204, 2007.

[50] Z. Abid, M. Liu, and W. Wang, “3D Integration of CMOL Structures for FPGA Applications,” *IEEE Trans. Comp.*, vol. 60, pp. 463-471, 2011.

[51] M. Liu and W. Wang, “Application of nanojunction-based RRAM to reconfigurable IC,” *Micro Nano Lett.*, vol. 3, pp. 101-105, 2008.

[52] S. Paul and S. Bhunia, “Computing with Nanoscale Memory: Model and Architecture,”, 2009, pp. 1-6.

[53] S. Paul and S. Bhunia, “A Scalable Memory-Based Reconfigurable Computing Framework for Nanoscale Crossbar,” *IEEE Trans. Nanotechnol.*, vol. 11, pp. 451-462, 2012.

[54] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, and R. S. Williams, “‘Memristive’ switches enable ‘stateful’ logic operations via material implication,” *Nature*, vol. 464, pp. 873-876, 2010.

[55] P. J. Kuekes, D. R. Stewart, and R. S. Williams, “The crossbar latch: logic value storage, restoration, and inversion in crossbar circuits,” *J. Appl. Phys.*, vol. 97, pp. 34301/1-5, 2005.

[56] G. Snider, “Computing with hysteretic resistor crossbars,” *Appl. Phys. A - Mater. Sci. Process.*, vol. A80, pp. 1165-1172, 2005.

[57] E. Linn, S. Menzel, S. Ferch, and R. Waser, “Compact modeling of CRS devices based on ECM cells for memory, logic and neuromorphic applications,” *Nanotechnology*, vol. 24, pp. 384008, 2013.

[58] ITRS-The International Technology Roadmap for Semiconductors, “Edition 2012,” *http://www.itrs.net/*, 2012.

[59] A. Chen, “A Comprehensive Crossbar Array Model With Solutions for Line Resistance and Nonlinear Device Characteristics,” *IEEE Trans. Electron Devices*, vol. 60, pp. 1318 - 1326, 2013.

[60] C. Schindler, G. Staikov, and R. Waser, “Electrode kinetics of Cu-SiO_2-based resistive switching cells: Overcoming the voltage-time dilemma of electrochemical metallization memories,” *Appl. Phys. Lett.*, vol. 94, pp. 072109, 2009.

[61] T. Breuer, A. Siemon, E. Linn, S. Menzel, R. Waser, and V. Rana, “A HfO_2-Based Complementary Switching Crossbar Adder,” *Advanced Electronic Materials*, DOI: 10.1002/aelm.201500138, 2015. [Early Online Access]

[62] A. Siemon, S. Menzel, A. Chattopadhyay, R. Waser, and E. Linn, “In-Memory Adder Functionality in 1S1R Arrays,” *ISCAS*, pp. 1338-1341, 2015.

[63] S. O. Haykin, *Neural Networks and Learning Machines* Pearson International Edition, 2009.

[64] S. Geman, E. Bienenstock, and R. Doursat, "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, vol. 4, pp. 1-58, 1992.

[65] W. Singer, "Binding by Synchrony," *Scholarpedia*, vol. 2, pp. 1657, 2007.

[66] L. Chittka and A. Brockmann, "Perception Space - The Final Frontier," *PLoS Biol*, vol. 3, pp. e137, 2005.

[67] W. Gerstner and W. Kistler, *Spiking Neuron Models* Cambridge University Press, Cambridge, 2002.

[68] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *Journal of Physiology*, vol. 117, pp. 500-544, 1952.

[69] C. Koch, *Biophysics of Computation: Information Processing in Single Neurons* Oxford University Press, Oxford, 2004.

[70] G.-Q. Bi and M.-M. Poo, "Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type," *Journal of Neuroscience*, vol. 18, pp. 10464-10472, 1998.

[71] H. Markram, Y. Wang, and M. Tsodyks, "Differential signaling via the same axon of neocortical pyramidal neurons," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 5323-5328, 1998.

[72] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, "Rate, Timing, and Cooperativity Jointly Determine Cortical Synaptic Plasticity," *Neuron*, vol. 32, pp. 1149-1164, 2001.

[73] U. Ramacher and C. v. d. Malsburg, *On the Construction of Artificial Brains* Springer, 2010.

[74] D. O. Hebb, *The Organization of Behavior* New York: Wiley & Sons, 1949.

[75] A. Heittmann and U. Ramacher, "Correlation-based Feature Detection Using Pulsed Neural Networks," *IEEE 13th Workshop on Neural Networks for Signal Processing (NNSP)*, pp. 479-488, 2003.

[76] J. Schreiter, U. Ramacher, A. Heittmann, D. Matolin, and R. Schüffny, "Cellular Pulse Coupled Neural Network with Adaptive Weights for Image Segmentation and its VLSI Implementation," *SPIE Proceedings*, vol. 5298, pp. 290-296, 2004.

[77] A. Heittmann and Ramacher, "Pulsed Neural Networks for Feature Detection Using Dynamic Synapses," *4th International ICSC Symposium on Engineering of Intelligent Systems (EIS)*, 2004.

[77a] Indiveri, G., Chicca, E., Douglas, R., "A VLSI Array of Low-Power Spiking Neurons and Bistable Synapses With Spike-Timing Dependent Plasticity*", IEEE Transactions on Neural Networks*, vol. 17, pp. 211-221, 2006

[78] T. Kohonen, *Self-organization and associative memory* Springer, 1989.

[79] G. Indiveri, B. Linares-Barranco, R. Legenstein, G. Deligeorgis, and T. Prodromakis, "Integration of nanoscale memristor synapses in neuromorphic computing architectures," *Nanotechnology*, vol. 24, pp. 384010, 2013.

[80] O. Turel and K. Likharev, "CrossNets: possible neuromorphic networks based on nanoscale components," *International Journal of Circuit Theory and Applications*, vol. 31, pp. 37-53, 2003.

[81] H. Choi, H. Jung, J. Lee, J. Yoon, J. Park, D. J. Seong, W. Lee, M. Hasan, G. Y. Jung, and H. Hwang, "An electrically modifiable synapse array of resistive switching memory," *Nanotechnology*, vol. 20, pp. 345201, 2009.

[82] F. Alibart, T. Sherwood, and D. Strukov, "Hybrid CMOS/Nanodevice circuits for high throughput pattern matching applications," *Proceedings of the 2011 Nasa/Esa Conference on Adaptive Hardware and Systems (AHS)*, pp. 279-286, 2011.

[83] A. Flocke, T.G. Noll, C. Kugeler, C. Nauenheim, and R. Waser, "A fundamental analysis of nano-crossbars with non-linear switching materials and its impact on TiO_2 as a resistive layer," *IEEE*, pp. 319-322, 2008.

[84] G.S. Snider, "Spike-timing-dependent learning in memristive nanodevices," *IEEE International Symposium on Nanoscale Architectures*, pp. 85 - 92, 2008.

[85] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale Memristor Device as Synapse in Neuromorphic Systems," *Nano Lett.*, vol. 10, pp. 1297-1301, 2010.

[86] Z. Wei, T. Takagi, Y. Kanzawa, Y. Katoh, T. Ninomiya, K. Kawai, S. Muraoka, S. Mitani, K. Katayama, S. Fujii, R. Miyanaga, Y. Kawashima, T. Mikawa, K. Shimakawa, and K. Aono, "Retention model for high-density ReRAM," *4th IEEE International Memory Workshop (IMW)*, 2012.

[87] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, and M. Aono, "Short-term plasticity and long-term potentiation mimicked in single inorganic synapses," *Nat. Mater.*, vol. 10, pp. 591-595, 2011.

[88] I. Valov, E. Linn, S. Tappertzhofen, S. Schmelzer, J. v. d. Hurk, F. Lentz, and R. Waser, "Nanobatteries in redox-based resistive switches require extension of memristor theory," *Nature Communications*, vol. 4, pp. 1771, 2013.

[89] S. Tappertzhofen, E. Linn, U. Böttger, R. Waser, and I. Valov, "Nanobattery Effect in RRAMs - Implications on Device Stability and Endurance," *IEEE Electron Device Lett.*, vol. 35, pp. 208-210, 2014.

[90] S. Menzel, S. Tappertzhofen, R. Waser, and I. Valov, "Switching Kinetics of Electrochemical Metallization Memory Cells," *PCCP*, vol. 15, pp. 6945-6952, 2013.

[91] D. Querlioz, O. Bichler, and C. Gamrat, "Simulation of a Memristor-Based Spiking Neural Network Immune to Device Variations," *International Joint Conference on Neural Networks (IJCNN)*, pp. 1775-1781, 2011.

[92] C. Zamarreño-Ramos, L. A. Camuñas-Mesa, J. A. Pérez-Carrasco, Timothée Masquelier, T. Serrano-Gotarredona, and B. Linares-Barranco, "On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex," *Frontiers in Neuroscience*, vol. 5, pp. 26, 2011.

[93] T. Serrano-Gotarredona, T. Masquelier, T. Prodromakis, G. Indiveri, and B. Linares-Barranco, "STDP and STDP variations with memristors for spiking neuromorphic learning systems," *Frontiers in Neuroscience*, vol. 7, pp. 2, 2013.

[94] U. Russo, D. Kamalanathan, D. Ielmini, A. L. Lacaita, and M. N. Kozicki, "Study of Multilevel Programming in Programmable Metallization Cell (PMC) Memory," *IEEE Trans. Electron Devices*, vol. 56, pp. 1040-1047, 2009.

[95] S. Menzel, U. Böttger, and R. Waser, "Simulation of multilevel switching in electrochemical metallization memory cells," *J. Appl. Phys.*, vol. 111, pp. 014501/1-5, 2012.

[96] S. Balatti, S. Larentis, D. C. Gilmer, and D. Ielmini, "Multiple Memory States in Resistive Switching Devices Through Controlled Size and Orientation of the Conductive Filament," *Advanced Materials*, vol. 25, pp. 1474-1478, 2013.

[97] S. Ambrogio, S. Balatti, F. Nardi, S. Facchinetti, and D. Ielmini, "Spike-timing dependent plasticity in a transistor-selected resistive switching memory." *Nanotechnology*, vol. 24, pp. 384012-384012, 2013.

[98] M. Suri, O. Bichler, D. Querlioz, and G. Palma, "CBRAM devices as binary synapses for low-power stochastic neuromorphic systems: Auditory (Cochlea) and visual (Retina) cognitive processing applications," *International Electron Devices Meeting (IEDM)*, pp. 10.3.1 - 10.3.4, 2012.

[99] S. Tappertzhofen, E. Linn, L. Nielen, R. Rosezin, F. Lentz, R. Bruchhaus, I. Valov, U. Böttger, and R. Waser, "Capacity based Nondestructive Readout for Complementary Resistive Switches," *Nanotechnology*, vol. 22, pp. 395203, 2011.

[100] O. Kavehei, E. Linn, L. Nielen, S. Tappertzhofen, S. Skafidas, I. Valov, and R. Waser, "Associative Capacitive Network based on Nanoscale Complementary Resistive Switches for Memory-Intensive Computing," *Nanoscale*, vol. 5, pp. 5119-5128, 2013.

The IFF Spring School in Jülich, the Peter Grünberg Institute, and JARA-FIT

PGI/JCNS-TA
52425 Jülich
Phone: ++49 2461 61-1739
Fax: ++49 2461 61-2410
Email: springschool@fz-juelich.de
web: www.iff-springschool.de

The annual IFF Spring Schools were first brought into being in 1970 by the Institut für Festkörperforschung (IFF). Since then, these schools have made it possible for students and young scientists to gain a two-week insight into a current topic related to condensed matter physics.

As a result of a restructuring of our research organization in 2011, four new institutes emerged from the former IFF and the former IBN (Institute for Bio- and Nanosystems): Research in electronic systems and phenomena, as well as their applications in nanoelectronics and information technology, is located in the Peter Grünberg Institut (PGI), named after our colleague who received the Nobel Prize for physics in 2007. The Jülich Centre for Neutron Science (JCNS) is dedicated to the operation of neutron scattering instruments and national and international neutron sources. Soft matter and biophysics research is integrated into the Institute of Complex Systems (ICS). The Institute for Advanced Simulation (IAS) focuses on developing and applying high-performance computing for quantum phenomena, solid-state research, and complex systems. The Spring School is organized in succession by these four institutes.

The PGI consists of eleven departments: Quantum Theory of Materials, Theoretical Nanoelectronics, Functional Nanostructures at Surfaces, Scattering Methods, Microstructure Research, Electronic Properties, Electronic Materials, Bioelectronics, Semiconductor Nanoelectronics, and the JARA Institutes for Green Information Technology and for Quantum Information. We operate the Helmholtz Nanoelectronics Facility (HNF) and, together with the Central Facility for Electron Microscopy at the RWTH Aachen University, the Ernst Ruska-Centre (ER-C) for Microscopy and Spectroscopy with Electrons. In addition, our departments participate in the operation of synchrotron and neutron beam lines as well as the Jülich supercomputers. We are part of the Jülich Aachen Research Alliance within the section Fundamentals of Future Information Technology (JARA-FIT) in which we collaborate with the physicists, chemists, electrical engineers, material scientists, and biologists of the RWTH Aachen University. In a concerted way, we conduct exploratory research in nanoelectronics and quantum phenomena with an emphasis on potential long-term applications in information technology and beyond.

Forschungszentrum Jülich – Campus Map

Band / Volume 113
Memristive Phenomena - From Fundamental Physics to Neuromorphic Computing
Lecture Notes of the 47th IFF Spring School **2016**
22 February - 04 March 2016
ed. by R. Waser and M. Wuttig
ISBN: 978-3-95806-091-3

Band / Volume 94
Functional Soft Matter
Lecture Notes of the 46th IFF Spring School **2015**
23 February - 06 March 2015
ed. by J. Dhont, G. Gompper, G. Meier, D. Richter, G. Vliegenthart and R. Zorn
ISBN: 978-3-89336-999-7

Band / Volume 74
Computing Solids - Models, ab-initio methods and supercomputing
Lecture Notes of the 45th IFF Spring School **2014**
10 - 21 March 2014
ed. by S. Blügel, N. Helbig, V. Meden and D. Wortmann
ISBN: 978-3-89336-912-6

Band / Volume 52
Quantum Information Processing
Lecture Notes of the 44th IFF Spring School **2013**
25 February - 08 March 2013
ed. by D. DiVincenzo
ISBN: 978-3-89336-833-4

Band / Volume 33
Scattering Methods for Condensed Matter Research: Towards Novel Applications at Future Sources
Lecture Notes of the 43rd IFF Spring School **2012**
05 - 16 March 2012
ed. by M. Angst, T. Brückel, D. Richter, R. Zorn
ISBN: 978-3-89336-759-7

Band / Volume 20
Macromolecular Systems in Soft- and Living-Matter
Lecture Notes of the 42nd IFF Spring School **2011**
14 – 25 February 2011
ed. by J. K.G. Dhont, G. Gompper, P. R.Lang, D. Richter, M. Ripoll, D. Willbold and R. Zorn
ISBN: 978-3-89336-688-0

Band / Volume 13
Electronic Oxides - Correlation Phenomena, Exotic Phases and Novel Functionalities
Lecture Notes of the 41st IFF Spring School **2010**
08 – 19 March 2010
ed. by S. Blügel, T. Brückel, R. Waser and C.M. Schneider
ISBN: 978-3-89336-609-5

Band / Volume 10
Spintronics – From GMR to Quantum Information
Lecture Notes of the 40th IFF Spring School **2009**
09 – 20 March 2009
ed. by S. Blügel, D. Bürgler, M. Morgenstern, C. M. Schneider and R. Waser
ISBN: 978-3-89336-559-3

Band / Volume 1
Soft Matter - From Synthetic to Biological Materials
Lecture Notes of the 39th IFF Spring School **2008**
03 – 14 March 2008
ed. by J.K.G. Dhont, G. Gompper, G. Nägele, D. Richter and R.G. Winkler
ISBN: 978-3-89336-517-3

Band / Volume 34
Probing the Nanoworld - Microscopies, Scattering and Spectroscopies of the Solid State
Lecture Notes of the 38th IFF Spring School **2007**
12 - 23 March 2007
ed. by K. Urban, C. M. Schneider, T. Brückel, S. Blügel, K. Tillmann, W. Schweika, M. Lentzen and L. Baumgarten
ISBN: 978-3-89336-462-6

Band / Volume 37
Computational Condensed Matter Physics
Lecture Notes of the 37th IFF Spring School **2006**
06 - 17 March 2006
ed. by S. Blügel, G. Gompper, E. Koch, H. Müller-Krumbhaar, R. Spatschek and R. G. Winkler
ISBN: 978-3-89336-430-5

Band / Volume 26
Magnetism goes Nano - Electron Correlations, Spin Transport, Molecular Magnetism
Lecture Notes of the 36th IFF Spring School **2005**
14 - 25 February 2005
ed. by S. Blügel, T. Brückel and C. M. Schneider
ISBN: 3-89336-381-5

Band / Volume 19
Physics meets Biology - From Soft Matter of Cell Biology
Lecture Notes of the 35th IFF Spring School **2004**
22 March - 02 April 2004
ed. by G. Gompper, U. B. Kaupp, J. K. G. Dhont, D. Richter and R. G. Winkler
ISBN: 3-89336-348-3

Band / Volume 14
Fundamentals of Nanoelectronics
Lecture Notes of the 34th IFF Spring School **2003**
10 - 21 March 2003
ed. by S. Blügel, M. Luysberg, K. Urban and R. Waser
ISBN: 3-89336-319-X

Band / Volume 10
Soft Matter - Complex Materials on Mesoscopic Scales
Lecture Notes of the 33rd IFF Spring School **2002**
04 - 15 March 2002
ed. by J.K.G. Dhont, G. Gompper and D. Richter
ISBN: 3-89336-297-5

Band / Volume 7
Neue Materialien für die Informationstechnik
Vorlesungsmanuskripte des 32. IFF-Ferienkurses **2001**
05. - 16. März 2001
R. Waser (Editor)
ISBN: 3-89336-279-7